

# CENTURY Academic Assessments

## Technical Manual

This manual covers curriculum-aligned standardised assessments in English (SPaG), Mathematics and Science for Years 2 to 10 in schools following the English National Curriculum.

|                       |                         |
|-----------------------|-------------------------|
| Version               | 1.0                     |
| Date of publication   | September 2026          |
| Status                | Published               |
| Supersedes            | First published version |
| Next scheduled review | Spring 2027             |

This manual sets out the methodology shared across every assessment in the suite. Per-assessment results are published in a Results Supplement for each assessment, which reports its item trial, standardisation sample, reliability, standard error of measurement and differential item functioning. Each Supplement cites this Manual by section number and states the Manual version it was written against.

How to cite: CENTURY Tech (2026). CENTURY Academic Assessments: Technical Manual, version 1.0. London: CENTURY Tech.

Enquiries: [assessments@century.tech](mailto:assessments@century.tech)

© CENTURY Tech Ltd 2026. All rights reserved.

## Authorship and contributions

| Role                                            | Name                                                                                                           |
|-------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| Document owner and lead author                  | Chris Richards,<br>Head of Curriculum and Assessment Development                                               |
| Blueprint development and item authoring        | CENTURY's Curriculum and Assessment<br>Development team                                                        |
| Subject lead, English                           | Victoria Tripp                                                                                                 |
| Subject lead, Mathematics                       | Karen Lea                                                                                                      |
| Subject lead, Primary                           | Tori Evans                                                                                                     |
| Subject lead, Science                           | Tom Feather                                                                                                    |
| Item trial and standardisation analysis         | Robyn Armstrong-Holding, using the analytical<br>framework and code provided by Select Statistical<br>Services |
| Statistical partner and psychometric<br>advisor | Select Statistical Services: Lynsey McColl<br>& Jo Morrison                                                    |

The division of responsibility between CENTURY and Select Statistical Services is set out in §2.1.

### Version history

| Version | Date           | Summary of changes      | Author         |
|---------|----------------|-------------------------|----------------|
| 1.0     | September 2026 | First published version | Chris Richards |

Section numbering is unchanged between versions 1.0 and any subsequent minor version. Where a major version changes section numbering, the affected Results Supplements are reissued with updated cross-references.

|                                                          |    |
|----------------------------------------------------------|----|
| Version history.....                                     | 1  |
| 1. Introduction.....                                     | 4  |
| 1.1 Purpose.....                                         | 4  |
| 1.2 The CENTURY Academic Assessments Suite.....          | 5  |
| 2. Development of the assessments.....                   | 7  |
| 2.1 Development overview.....                            | 7  |
| Owners2.2 Blueprint development.....                     | 8  |
| 2.3 Item authoring.....                                  | 10 |
| 2.4 Internal review and proofing.....                    | 12 |
| 2.5 Internal user testing.....                           | 12 |
| 2.6 Item trial methodology.....                          | 13 |
| 2.7 Standardisation methodology.....                     | 17 |
| 3. Reliability, SEM and DIF.....                         | 19 |
| 3.1 Reliability.....                                     | 19 |
| 3.2 Standard Error of Measurement (SEM).....             | 20 |
| 3.3 Confidence intervals around standardised scores..... | 20 |
| 3.4 Differential Item Functioning (DIF).....             | 21 |
| 4. Validity.....                                         | 23 |
| 4.1 Approach to validity.....                            | 23 |
| 4.2 Content validity.....                                | 23 |
| 4.3 Internal structure.....                              | 24 |
| 4.4 Fairness.....                                        | 25 |
| 4.5 Ongoing validity evidence.....                       | 25 |
| 5. Score reporting and interpretation.....               | 26 |
| 5.1 Reports in the CENTURY ecosystem.....                | 26 |
| 5.2 Interpreting individual scores.....                  | 27 |
| 5.3 Interpreting cohort results.....                     | 27 |
| 5.4 Interpreting progress over time.....                 | 28 |
| 5.5 Why standardised, not age-standardised.....          | 29 |
| 5.6 Common use cases.....                                | 30 |
| 5.7 Limits of interpretation.....                        | 30 |
| 6. Limitations and ongoing development.....              | 31 |
| 6.1 Suite development status.....                        | 31 |
| 6.2 Sample composition.....                              | 32 |
| 6.3 Post-pandemic standardisation.....                   | 33 |
| 6.4 Re-standardisation and ongoing review.....           | 33 |
| 6.5 Future development.....                              | 34 |
| 7. Glossary and references.....                          | 34 |
| 7.1 Glossary.....                                        | 34 |
| 7.2 References.....                                      | 37 |

Assessment Supplements..... 39

    Year 5 Baseline SPaG.....39

    Year 5 Baseline Mathematics.....43

    Year 5 Baseline Science..... 47

    Year 6 Baseline SPaG..... 51

    Year 6 Baseline Mathematics.....55

    Year 6 Baseline Science.....59

## 1. Introduction

### 1.1 Purpose

The CENTURY Academic Assessments are a suite of curriculum-aligned, standardised assessments in English (Spelling, Punctuation and Grammar, hereafter SPaG), Mathematics and Science. They are currently available for students in Years 2 to 10 in schools following the English National Curriculum. Available assessments can be found in Table 1.

### Construct measured

Each assessment is designed to measure a student's attainment against the relevant year group's expectations in the English National Curriculum for the named subject.

### Intended use

Each assessment is available in two parallel forms, which can be used on their own or together:

- a Baseline assessment, administered at the start of an academic year to establish students' starting points, identify retention gaps from the previous year and inform planning;
- an End-of-Year assessment, administered at the end of the academic year to summarise attainment over the year, support end-of-year reporting and inform transition to the next year group.

Each assessment is designed to be completed in a single 50-minute sitting under standard test conditions, lending itself to use within a normal lesson time period.

### Intended users and decisions

The assessments are intended to support teachers, subject leaders and senior leaders in:

- establishing each student's attainment relative to age-related national curriculum expectations and to a nationally representative sample;
- identifying strengths and areas for development at student, class and year-group level;
- informing curriculum planning, in-class teaching and targeted intervention;
- tracking progress over time, both within an academic year (between Baseline and End-of-Year sittings) and between academic years through comparable standardised scores referenced to the same national sample;

- providing consistent, comparable evidence for school self-evaluation and reporting.

## Outputs

Results are reported as raw scores, standardised scores (mean 100, standard deviation 15) and stanines (1 to 9), alongside curriculum-domain breakdowns, question-level analysis, comparison with the national average, and personalised follow-up resources within the CENTURY platform. Reports are produced at student and cohort levels.

## Use cautions

Results from the CENTURY Academic Assessments should be considered alongside other classroom evidence when teachers and schools draw conclusions about individual students or cohorts. They are not designed to be used as a substitute for national statutory assessments (for example, KS2 SATs, the Year 4 Multiplication Tables Check, or the KS1/KS2 Phonics Screening Check).

---

## 1.2 The CENTURY Academic Assessments Suite

### Delivery

Each assessment in the suite is administered online via the CENTURY platform. Schools schedule assessment windows in the Assessment Manager, students complete the assessment under standard test conditions, and items are auto-marked. Student and cohort reports are generated immediately and made available to teachers, with follow-up learning resources linked to the curriculum domains identified as areas for development.

### Subjects

Three subjects are assessed:

- **English – Spelling, Punctuation and Grammar (SPaG)**, aligned to the year group's English National Curriculum statements for SPaG and modelled on the format of the corresponding statutory test for the Key Stage;
- **Mathematics**, covering the year group's National Curriculum content across the topics in proportions appropriate to the year group;
- **Science**, covering the Biology, Chemistry and Physics strands of the National Curriculum.

The curriculum domains covered by each assessment are listed in the relevant Results Supplement.

### Year groups and assessment occasions

The suite spans Years 2 to 10 and is structured around two assessment occasions per year:

- a Baseline assessment at the start of the academic year (autumn term);
- an End-of-Year assessment at the end of the academic year (summer term).

**Primary:** Year 2 End-of-Year and Year 3 Baseline are available for Mathematics and SPaG only, with Science first available in the Year 3 End-of-Year; for Years 4 to 6, both Baseline and End-of-Year assessments are available across all three subjects.

**Secondary:** Year 7 is currently covered by Baseline in Mathematics and Science only; Mathematics baselines are also available in Year 8, 9 and 10. End-of-Year 7, 8 and 9 is available for Mathematics, with the End-of-Year 8 and 9 Science versions available depending on when a cohort completes their KS3 learning covering KS3 content.

Table 1 summarises the assessments currently available.

**Table 1: Assessments available in the suite**

| Year group | Baseline (autumn term) | End-of-Year (summer term) |
|------------|------------------------|---------------------------|
| Year 2     | —                      | Maths, SPaG               |
| Year 3     | Maths, SPaG            | Maths, SPaG, Science      |
| Year 4     | Maths, SPaG, Science   | Maths, SPaG, Science      |
| Year 5     | Maths, SPaG, Science   | Maths, SPaG, Science      |
| Year 6     | Maths, SPaG, Science   | Maths, SPaG, Science      |
| Year 7     | Maths, Science         | Maths                     |
| Year 8     | Maths                  | Maths, Science            |
| Year 9     | Maths                  | Maths, Science            |
| Year 10    | Maths                  |                           |

Availability is set out above. The development stage each assessment has reached, and the technical evidence published for it, are set out in §6.1.

## Parallel-form design

Where a Year  $N$  End-of-Year assessment has a corresponding Year  $N+1$  Baseline assessment in the same subject, the two are constructed as equated parallel forms: built from a common item pool, targeting the same construct, calibrated to the same difficulty, and balanced across the same curriculum domains, but differing in the specific items they contain.

This design allows a student to complete the End-of-Year assessment in the summer term and the Year  $N+1$  Baseline assessment the following autumn term without exposure to the same items. Both are reported on a common standardised scale, allowing schools to use both assessments in sequence to measure *summer learning loss* (reduction in retrieval strength) if desired, or may use only one (typically the End-of-Year) as their main summative measure.

More broadly, because all assessments in the suite are referenced to the same standardised score scale, student progress can be measured by comparing standardised scores between any two sittings, whether within an academic year or across years (see §5.4).

## Duration and format

Each assessment is designed to be completed in a single 50-minute sitting, fitting comfortably within a normal lesson. The total number of items per assessment varies by year group so that each assessment takes a similar length of time; younger students complete fewer items, reflecting realistic per-question times for the cohort. Items use a mix of formats consistent with the wider CENTURY platform, including multiple-choice, exact-answer, labelling and matching. Each item is worth one mark towards the raw score.

## 2. Development of the assessments

### 2.1 Development overview

The CENTURY Academic Assessments are developed through a multi-stage process. The process starts with the development of a curriculum-aligned blueprint, moves through item authoring and an item trial in which a larger pool of candidate items is administered, applies item analysis to identify items that perform well statistically, selects items into final assessments, and concludes with a standardisation trial with a representative sample of schools that produces the score conversions used in live reporting.

The CENTURY Academic Assessments are developed and owned by CENTURY. Select Statistical Services (hereafter Select) is CENTURY's statistical partner and psychometric advisor for the suite, providing the analytical framework and code underpinning the item analysis and standardisation described in this manual. CENTURY is responsible for curriculum mapping, item authoring, sample recruitment, test administration, item selection and the ongoing publication of the assessments.

The detailed methodology for each stage is set out in §3 to §6 of this manual.

The stages are:

1. **Blueprint development:** a blueprint is created for each assessment, specifying the curriculum domains to be covered, the proportion of items per domain, the target difficulty distribution, and the item types to be used.
2. **Item authoring:** items are written in line with the curriculum blueprint for each subject and year group.
3. **Item trial:** candidate items are administered in a spiralled design across multiple parallel forms to a sample of schools to estimate each item's statistical characteristics.
4. **Item analysis:** classical and IRT statistics are calculated for each item.
5. **Further item authoring:** additional items are written to fill gaps in difficulty coverage or curriculum representation identified by the analysis.
6. **Test construction:** final assessment forms are assembled from retained items, balanced for difficulty, domain coverage and reliability.
7. **Test analysis:** the constructed forms are evaluated for overall reliability and test-level statistical properties before standardisation.
8. **Standardisation trial:** the final forms are administered to a nationally representative sample to derive raw-to-standardised score conversions.
9. **Release:** the standardised assessment is made available for live use, with reporting against the national standardisation cohort.

Each stage is described in detail in the sections that follow.

## Owners2.2 Blueprint development

The blueprint is the foundational design document for each assessment. It is created before any items are authored and specifies the structure, content, balance and trial design of the assessment. A separate blueprint is produced for each subject within each year-pair (e.g. a single blueprint covers both the Year 4 End-of-Year and Year 5 Baseline Mathematics assessments).

Each blueprint specifies:

- **Live assessment size:** the number of items to appear in the final assessment, set in line with what students of the year group can reasonably complete within the 50-minute time limit.
- **Curriculum domain weightings:** the proportion of items allocated to each curriculum domain within the subject. Weightings are derived either from the relative number of statements within the year group's National Curriculum Programme of Study or from the published Key Stage statutory test framework percentage ranges, whichever is more appropriate to the subject.
- **National Curriculum mapping:** each domain is linked to specific National Curriculum Programme of Study statements for the year group.
- **CENTURY nugget mapping:** each Programme of Study statement is also linked to the corresponding CENTURY learning nugget, so that students who answer incorrectly can be directed to targeted learning content on the platform.
- **Complexity profile:** the target distribution of items across cognitive demand (Retrieval, Describe, Interpret, Demonstrate, Explain, Evaluate, Multi-skill), contextual demand, computational demand for Mathematics items, statistical interpretation demand for items involving graphs or tables, and response strategy (Selected Response, Short Constructed Response, Extended Constructed Response).
- **Trial design:** the number of unique trial items required, including an attrition allowance above the live target, the number of trial forms, and the spiralled rotation that places each item into forms. The trial design is described in full in §2.6.
- **Authoring assignments:** items within each domain are allocated to experienced subject specialists from CENTURY's curriculum team, with an independent reviewer assigned to formatting QA on each batch.

All blueprints reference a common **item style guide** that standardises command words, stem length, emboldening, image and graph conventions, exact-answer algorithm selection, item naming and visual style across the suite. The style guide is described in §2.3.

### Ownership and sign-off

Blueprints are owned and developed in-house by CENTURY's curriculum team, who hold the subject expertise for the assessments. Each blueprint is quality assured internally before being signed off by the relevant subject lead. Once signed off, the blueprint forms the basis for item authoring (§3.2), internal review (§3.3) and trial form construction (§4).

## 2.3 Item authoring

Items are authored by experienced subject specialists from CENTURY's curriculum team against the relevant blueprint (§2.2). Each item is written to align with a specific National Curriculum Programme of Study statement and a corresponding CENTURY learning nugget, and is targeted at a particular cell of the complexity profile (cognitive demand × contextual demand × response strategy).

### Item formats

Items use four core formats, consistent with those available across the wider CENTURY platform:

- **Multiple-choice.** A single correct answer is selected from a fixed set of options. Distractors are designed to reflect common misconceptions where possible.
- **Exact-answer.** The student types their answer into a response box. The marking algorithm is selected per item from:
  - **Numeric:** numerical input only, used where the answer is a number (e.g. an arithmetic calculation);
  - **Phonetic:** spelling tolerance applied, used where the answer is a word or short phrase and spelling is not itself the construct being assessed;
  - **Exact:** strict spelling required, used only where spelling is itself the construct being assessed (predominantly in SPaG items).
- **Labelling.** The student attaches a set of provided labels to specified points on a diagram or image.
- **Matching.** The student pairs items. Matching items are constrained in size so the item is viewable on a single screen,

All items are scored dichotomously: one mark for a fully correct response, zero otherwise. Each item contributes equally to the raw score.

### Item style guide

A common item style guide is applied across the suite to ensure consistency of presentation and to control variance in difficulty introduced by question wording or layout. Key style requirements include:

- **Command words:** items use plain interrogatives (*what, why, when, how, where, who*) where possible, combined sparingly with traditional command words (e.g. *describe, explain*).
- **Stem length:** capped, excluding any contextual setup.
- **Context:** items default to a Simple Familiar Context unless a higher cognitive demand justifies a novel or more complex context.
- **Emphasis:** bold is used consistently to highlight important information within the stem; italics are reserved for text extracts, binomial nomenclature and formal titles.
- **Layout:** question and instructions appear on separate lines; bullets are used to display strands of information clearly; answer lines are represented as five underscores.
- **Images and diagrams:** images are usually presented above the stem of the question.
- **Tables, graphs and figures:** graphs are restricted to a limited information density, use simpler gridlines where precise reading is not required, are black-and-white only, and have axes extended beyond the plotted data for clarity.
- **Item naming:** items are named by skill plus content (e.g. *Recall Photosynthesis*) so the item name is informative within the teacher-facing question-level analysis report.
- **Nugget linking:** every item is linked to the CENTURY learning nugget covering the corresponding curriculum content, enabling personalised follow-up resources after the assessment.

The full style guide is maintained in the master blueprint template and applied to every assessment in the suite.

### Accessibility

Items are authored to be accessible to the full range of students within the target year group:

- **Vocabulary** is kept in line with reading expectations typical for the year group. Technical or subject-specific terms are introduced only where they form part of the curriculum content being assessed.
- **Cultural capital** requirements outside the construct are consciously avoided. An item should be accessible on the basis of the curriculum content it is testing, not on the basis of prior knowledge of specific cultural references.

- **Colour** is checked for contrast against accessibility guidelines wherever it is used for a functional purpose (for example, to distinguish data series on a graph).

## 2.4 Internal review and proofing

Items are authored by experienced subject specialists from CENTURY's curriculum team and reviewed by a different specialist within the team or by the relevant subject lead. The reviewer is either another subject specialist within the team or the relevant subject lead. Items receive an additional check by the English team for SPaG, in addition to the standard curriculum-team review.

### Review pass on the drafted items

The first review takes place once items have been drafted, typically in a working document or spreadsheet alongside the blueprint. At this pass, the reviewer checks each item against:

- **Content correctness:** the item is technically and curricularly correct, and any single best answer is unambiguous.
- **Distractor quality:** distractors in multiple-choice items are plausible and reflect common student misconceptions, with no clueing or partial-credit shortcuts.
- **Blueprint alignment:** the item targets the National Curriculum statement, nugget, and complexity profile intended by the blueprint.
- **Style guide compliance:** the item conforms to the style requirements set out in §2.3 (command words, stem length, context, emphasis, layout, image and table conventions, naming).
- **Accessibility:** the item meets the accessibility expectations set out in §2.3 (vocabulary, cultural capital, colour contrast).

### Tracking and resolution

The QA status of every individual item is tracked through both review passes. Items flagged for discussion are revised and re-reviewed, or replaced. Items that are not flagged are either confirmed as ready to proceed or have already been corrected and re-checked. Only items with a clear pass at both review stages enter the item trial.

## 2.5 Internal user testing

Before items are released to the formal item trial, they are tested internally within CENTURY (by colleagues from teams outside the curriculum team that authored them). Many of these

reviewers are former teachers with classroom experience. The purpose of this pass is not statistical calibration – that is the role of the item trial (§2.6) – but a final user-experience check: that items render and behave as expected on the CENTURY platform, that instructions are intuitive, and that no last-minute ambiguities have slipped through earlier reviews.

Items that surface issues at this stage are revised and re-checked before being released to the trial sample.

## 2.6 Item trial methodology

### 2.6.1 Spiralled trial design

The item trial uses a **spiralled trial design**, in which items rotate across multiple forms so that every item is administered to a different subset of the trial sample. The specific arrangement used in this suite is sometimes referred to informally as a cartwheel design, after the visual pattern formed when the blocks are drawn in a circle.

The design has three benefits:

- The difficulty of every trial item can be estimated relative to every other item, because the overlapping forms link all items onto a common scale.
- Every item functions as an anchor item between forms, maximising the linkage across the trial.
- Item position is rotated across forms, balancing any systematic effects of fatigue or speededness across the trial.

### 2.6.2 Trial sample and administration

Schools are recruited to participate in the item trial through CENTURY's existing partner networks. Each participating school administers one or more of the six trial forms; no individual student sits more than one form. Trial samples are recruited to provide sufficient statistical power to estimate item characteristics under the spiralled design, in which each item is administered in two of the six forms.

Representativeness of the trial sample is considered during recruitment across student demographics including sex, eligibility for the Pupil Premium grant, English as an additional language (EAL), special educational needs (SEN), school location, and school sector. However, full national representativeness is not a hard requirement at the item-trial stage. The purpose of the trial is to estimate item-level statistical characteristics, not to define a normative score scale, and the relatively small per-item sample sizes mean that trial samples are unlikely to mirror the

national population exactly. National representativeness is enforced at the standardisation stage (§2.7).

Trial sittings are administered online under standard test conditions on the CENTURY platform, matching the conditions of the live assessment. Each student completes a single trial form within a single 50-minute sitting. Items are auto-marked through the platform, and item-level response data are extracted for analysis.

The trial dates, total sample sizes, per-form sample sizes, and the demographic composition of the achieved sample for each assessment are reported in the relevant Results Supplement.

### 2.6.3 Item analysis

The aim of the item analysis is to estimate each item's statistical characteristics, principally its difficulty and how well it discriminates between higher and lower attaining students.

For each subject and year-pair, classical item statistics<sup>1</sup> and Item Response Theory (IRT)<sup>2</sup> parameters from a two-parameter logistic (2-PL) model are calculated for all trialled items, with all six trial forms analysed together. The following statistics are collated:

- **Facility value:** the percentage of students who answered the item correctly.
- **Omission rate:** the percentage of students who skipped the item.
- **Not-reached rate:** the percentage of students who did not reach the item before time expired.
- **Corrected total-item correlation:** the classical item discrimination index.
- **Cronbach's alpha if item deleted:** the change in internal-consistency reliability that would result from removing this item from the item set.
- **IRT slope (a):** the item discrimination parameter under the 2-PL model.
- **IRT difficulty (b):** the item location parameter under the 2-PL model.
- **IRT expected raw score:** the expected score on the item for a student of average ability ( $\theta = 0$ ).

---

<sup>1</sup> The itemAnalysis() function in the CTT package was amended to handle pairwise missing data for the spiralled trial design. Willse JT (2018). *CTT: Classical Test Theory Functions*. R package version 2.3.3. doi:10.32614/CRAN.package.CTT. <https://CRAN.R-project.org/package=CTT>

<sup>2</sup> R. Philip Chalmers (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29. doi:10.18637/jss.v048.i06

IRT item characteristic curves (ICCs) are plotted for each item, overlaid with the raw response data grouped in deciles of total score, to provide a visual assessment of item fit and target difficulty.

All item statistics are collated into a working spreadsheet, alongside dynamic calculations of the expected raw score for the overall assessment and per domain, to support item selection (§2.6.4).

#### 2.6.4 Item selection

Items are selected from the trial pool for inclusion in the live assessment on the basis of the statistics calculated in §2.6.3. CENTURY's curriculum team applies a defined set of discard rules, supplemented by qualitative judgement on curriculum coverage and balance.

##### Discard rules

An item is discarded from the pool if **any** of the following apply:

- **Facility is outside 10% to 90%:** the item is either too easy or too hard to provide useful discrimination at the relevant ability range.
- **Omit-of-seen rate is 15% or higher:** students who reached the item are choosing not to attempt it, indicating an accessibility or presentation problem.
- **IRT slope is 0.4 or lower:** the item discriminates poorly between higher-attaining and lower-attaining students.
- **IRT difficulty parameter outside the range -3 to +3:** the item sits at an extreme of the latent ability scale, outside the range over which calibration is reliable.

##### Curriculum and balance constraints

In addition to the statistical screen, item selection must produce a final assessment that:

- meets the blueprint's curriculum-domain weightings (§2.2);
- covers the planned range of cognitive demand and response strategies;
- contains a sensible spread of difficulty across the item set, with sufficient items at all parts of the target ability range.

### 2.6.5 Final form construction and reliability check

Once items have been selected, the final assessment forms are constructed from the retained item pool. For each year-pair, two parallel forms are assembled: the Year  $N$  End-of-Year form and the Year  $N+1$  Baseline form (see §1.2). The two forms share no items in common, in line with the parallel-form design. Items are distributed between the two forms to balance:

- curriculum domain coverage, in line with the blueprint weightings;
- mean IRT-estimated difficulty across each form's items, so that the two forms target the same construct at the same average difficulty;
- expected raw scores, so that the two parallel forms yield equivalent scores for the same student;
- the standard deviation of IRT difficulty within each form, made as equivalent as possible between the two forms and made as large as possible within each form;
- the spread of items across complexity-profile cells;
- the location and span of the test information function across the ability range.

Maximising the spread of item difficulty within each form ensures that each assessment discriminates effectively across the full range of student ability, rather than only at a narrow band around the mean. A wider spread of item difficulty extends the test information function across the ability range, improves measurement precision for students at the extremes of the distribution, and reduces floor and ceiling effects in the raw score distribution. Matching the standard deviation between the two parallel forms ensures that the two forms provide equivalent measurement precision at all points on the ability range.

A final psychometric check is then performed by re-running the 2-PL IRT models on the items in each constructed form. The check produces:

- the **test information function** for each form, indicating the precision of the standardised score at each point on the ability range;
- the **marginal IRT reliability** for each form, summarising the consistency of scores across the full ability range.

The target for marginal IRT reliability of each final form is at least 0.85. Forms that do not meet this threshold are restructured by reallocating items between the two forms before the assessment proceeds to the standardisation trial (§2.7).

The marginal IRT reliabilities and test information functions achieved for each form are reported in the relevant Results Supplement.

## 2.7 Standardisation methodology

### 2.7.1 Standardisation trial

Once an assessment has been constructed (§2.6.5), it is administered in a standardisation trial to produce the raw-to-standardised score conversions used in live reporting.

**Sample target.** The standardisation trial aims for at least 1,000 students per assessment, recruited to be broadly representative of the national population across student demographics including sex, eligibility for the Pupil Premium grant, English as an additional language (EAL), special educational needs (SEN), school location, and school sector.

**Timing.** The standardisation trial takes place in the next suitable period after the item trial. Each assessment is standardised in the academic term in which it is intended to be used: End-of-Year assessments are standardised during the summer term of the relevant year group, and Baseline assessments are standardised during the early autumn term for the relevant year group.

**Administration.** Standardisation sittings are administered under standard test conditions on the CENTURY platform, matching the conditions of the final assessment. Each student completes the full assessment within a single 50-minute sitting, and items are auto-marked through the platform.

### 2.7.2 Sample composition and weighting

Demographic data on participating students are collected from schools at the point of enrolment to the standardisation trial. The achieved sample composition is then compared with national population figures published by the Department for Education<sup>3,4</sup> using one-sample chi-square tests, across the dimensions of student sex, Pupil Premium eligibility, EAL, SEN, school location, and school sector.

Where the achieved sample deviates significantly from the national population on a dimension that is associated with attainment, sample weights are applied during the standardisation calculations so that the effective sample reflects the national population on that dimension. The specific dimensions that are weighted vary between individual assessments, depending on the cohort composition achieved for each trial.

---

<sup>3</sup> <https://explore-education-statistics.service.gov.uk/>

<sup>4</sup> <https://get-information-schools.service.gov.uk/Downloads>

The sample composition, chi-square comparisons against the national population, and the dimensions on which weights were applied for each assessment are reported in the relevant Results Supplement.

### 2.7.3 Raw-to-standardised score conversion

Raw scores from the standardisation sample are converted to standardised scores with a mean of 100 and a standard deviation of 15 by fitting a model using Taylor polynomials<sup>5</sup>.

No age effect is included in the conversion: each student is compared against the cohort sitting the assessment at the same point in the academic year, against the same year-group expectations, regardless of within-year age variation.

### 2.7.4 Stanines

Standardised scores are also reported as stanines on a 1-to-9 scale. Stanines provide a coarser-grained band classification that supports quick interpretation at the classroom level. Each stanine corresponds to a defined range of standardised scores, derived from cut-points at intervals of half a standard deviation on the standardised scale. The conversion is shown in Table 2.

**Table 2: Standardised score to stanine conversion**

| Stanine | Standardised score range | Approximate population proportion |
|---------|--------------------------|-----------------------------------|
| 1       | < 74                     | 4%                                |
| 2       | 74 – 81                  | 7%                                |
| 3       | 82 – 88                  | 12%                               |
| 4       | 89 – 96                  | 17%                               |
| 5       | 97 – 103                 | 20%                               |
| 6       | 104 – 111                | 17%                               |
| 7       | 112 – 118                | 12%                               |
| 8       | 119 – 126                | 7%                                |
| 9       | ≥ 127                    | 4%                                |

<sup>5</sup> Lenhard, A., Lenhard, W., Gary, S. (2018). *Continuous Norming (cNORM)*. The Comprehensive R Network, Package cNORM. <https://CRAN.R-project.org/package=cNORM>

### 2.7.5 Equating across parallel forms

The two parallel forms of each year-pair (the Year  $N$  End-of-Year form and the Year  $N+1$  Baseline form) are each standardised independently against their own normative cohort, at the timing for which they are intended to be used. The Year  $N$  End-of-Year form is standardised against the cohort sitting it in the summer term of Year  $N$ ; the Year  $N+1$  Baseline form is standardised against the cohort sitting it in the early autumn term of Year  $N+1$ .

Both forms are normed to the same standardised-score scale (mean 100, standard deviation 15) and the same stanine cut-points. A student's standardised score on either form therefore represents their position relative to the appropriate cohort against the year-group expectations at that point in the academic year. Because both forms share a common scale and a common interpretive frame, scores are directly comparable across the two forms, supporting measurement of progress between summer and autumn-term sittings, and across academic years.

## 3. Reliability, SEM and DIF

This section documents the reliability evidence and the fairness evidence that supports the use of the standardised scores reported by the CENTURY Academic Assessments.

Reliability is reported in two complementary ways:

- the **IRT marginal reliability** of each constructed form, calculated during item selection and final form construction (described in §2.6.5);
- the **Cronbach's alpha** internal-consistency reliability of each assessment, calculated on its standardisation sample (described in §3.1).

The **standard error of measurement (SEM)** is derived from these reliabilities and used to construct confidence intervals around individual student scores (§3.2 and §3.3).

In addition, every assessment is examined for **differential item functioning (DIF)** to identify items that may function differently for different groups of students (§3.4).

The reliability, SEM, confidence interval and DIF values achieved for each assessment are reported in the relevant Results Supplement.

### 3.1 Reliability

Cronbach's alpha is calculated for each assessment on its standardisation sample, providing a classical internal-consistency reliability estimate that complements the IRT marginal reliability calculated during final form construction (§2.6.5).

Cronbach's alpha measures the extent to which items within an assessment provide consistent measurement of the same underlying construct. Values range between 0 and 1, with higher values indicating greater internal consistency. Conventionally, Cronbach's alpha of at least 0.80 is considered acceptable for assessments used in low-stakes educational decision-making, and at least 0.90 for assessments used in higher-stakes individual decision-making.

The target for Cronbach's alpha on each CENTURY Academic Assessment is at least 0.80, and for marginal IRT reliability at least 0.85 (§2.6.5). Where an assessment records a value below its target, the value is reported in the relevant Results Supplement together with the reason, the effect on the standard error of measurement and the confidence interval, plus the action being taken, which will normally be to review the item at the next re-standardisation (§6.4). No assessment is withheld from reporting because of a below-target reliability, and the target is not adjusted to accommodate an individual result.

### 3.2 Standard Error of Measurement (SEM)

A student's observed score on an assessment reflects both their true level of ability on the construct measured and a degree of measurement error introduced by, for example, the particular items in the form, the conditions on the day, and the inherent uncertainty of any psychometric measurement.

The **standard error of measurement (SEM)** quantifies this measurement uncertainty. It is calculated as:

$$SEM = SD \times \sqrt{1 - \alpha}$$

where SD is the standard deviation of the standardised score distribution (15 by definition) and  $\alpha$  is the Cronbach's alpha reliability of the assessment.

The SEM is reported in standardised score points and indicates the typical magnitude of measurement error around a single score. A smaller SEM indicates higher precision of measurement; a larger SEM indicates that scores should be treated with greater caution.

The SEM values for each assessment are reported in the relevant Results Supplement.

### 3.3 Confidence intervals around standardised scores

The SEM is used to construct **confidence intervals** around individual standardised scores. A confidence interval is the range within which a student's true level of ability on the construct is likely to fall, taking into account the measurement error in their observed score.

Two confidence levels are commonly used:

- A **68% confidence interval** spans one SEM either side of the observed score. If the student were to sit comparable assessments many times, around 68 in every 100 observed scores would fall within this range.
- A **95% confidence interval** spans approximately two SEMs either side of the observed score (more precisely, 1.96 SEMs). If the student were to sit comparable assessments many times, around 95 in every 100 observed scores would fall within this range.

The 95% confidence interval is the conventional level for reporting in educational measurement and is recommended when interpreting an individual student's score.

#### Worked example

Suppose an assessment in the suite has an SEM of 5 standardised score points. For a student with an observed standardised score of 100:

- the **95% confidence interval** is  $100 \pm (2 \times 5) = \mathbf{90 \text{ to } 110}$ ;
- the **68% confidence interval** is  $100 \pm 5 = \mathbf{95 \text{ to } 105}$ .

In practical terms, while the student's best-estimate standardised score is 100, repeated measurement of their ability would produce a range of scores: around 95 in every 100 of those scores would fall between 90 and 110. The interval therefore reflects the inherent uncertainty in any single measurement. Two students with observed standardised scores of, say, 98 and 102 should not be interpreted as differing meaningfully in attainment, because their 95% confidence intervals overlap substantially. Conversely, two students with observed scores of 90 and 110 (four SEMs apart) can be interpreted with confidence as having genuinely different levels of attainment.

The SEM and worked confidence interval values for each assessment are reported in the relevant Results Supplement.

### 3.4 Differential Item Functioning (DIF)

Differential item functioning (DIF) analysis is carried out on the standardisation sample for each assessment to identify items that may function differently for different groups of students.

#### What DIF measures

DIF occurs where students from one group (for example, female students) have a higher or lower probability of answering an item correctly compared to students of the same ability in another group (for example, male students). DIF is distinct from group-level differences in mean

performance: a DIF analysis controls for ability, and tests whether the item itself behaves differently across groups.

### Groups examined

DIF is analysed by sex and by eligibility for the Pupil Premium grant. These are the dimensions on which individual-level data of sufficient quality are currently collected in the standardisation samples. The method applies to any demographic dimension for which such data are available; extending the analysis to English as an additional language and to special educational needs is set out in §6.5.

### Statistical method

Items are tested for DIF using a 2-parameter logistic (2-PL) multiple-group IRT model. The p-values from the multiple-group tests are adjusted for multiple testing using the Benjamini-Hochberg procedure<sup>6</sup> to control the false discovery rate.

### Classifying the magnitude of DIF

Statistical significance of DIF does not necessarily indicate a problem. Items showing statistically significant DIF may have very small effect sizes that are not meaningful in practice, and items showing DIF in opposite directions across an assessment may cancel out at the assessment level so that the assessment as a whole is balanced.

The magnitude of DIF is classified using the scale developed by Jodoin and Gierl (2001)<sup>7</sup> as negligible, moderate or large.

### Action policy

CENTURY's response to DIF identified during analysis is set out in Table 3.

**Table 3: Action taken in response to DIF findings**

| DIF magnitude | Action                                                                                                                                         |
|---------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| Negligible    | No action; item retained.                                                                                                                      |
| Moderate      | A subject specialist reviews the item for a plausible source of bias; the item is edited, replaced, or retained with documented justification. |

<sup>6</sup> Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

<sup>7</sup> Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, 14(4), 329–349. [https://doi.org/10.1207/S15324818AME1404\\_2](https://doi.org/10.1207/S15324818AME1404_2)

---

|       |                                                                                                                 |
|-------|-----------------------------------------------------------------------------------------------------------------|
| Large | The item is removed from the assessment in the next available revision; a replacement is authored and trialled. |
|-------|-----------------------------------------------------------------------------------------------------------------|

---

For each assessment, the number of items showing statistically significant DIF on each dimension examined, the magnitude classification of those items, and any subsequent action taken are reported in the relevant Results Supplement.

## 4. Validity

### 4.1 Approach to validity

Validity refers to the degree to which evidence and theory support the interpretations of standardised scores from an assessment for its intended uses. Validity is a property of the uses of an assessment's scores, not of the assessment itself.

The CENTURY Academic Assessments are designed to support the uses set out in §1.1: identifying student attainment relative to age-related national curriculum expectations and to a nationally representative sample, identifying strengths and areas for development at student and cohort level, informing curriculum planning and intervention, tracking progress over time, and providing comparable evidence for school self-evaluation and reporting.

This section documents the validity evidence currently available for those uses, presented under three headings:

- **Content validity** (§4.2): the extent to which the assessments cover the construct they are designed to measure.
- **Internal structure** (§4.3): the extent to which the items within an assessment work together to measure a single, coherent construct.
- **Fairness** (§4.4): the extent to which the assessments measure the construct equally well across different groups of students.

Validity evidence is cumulative. The evidence below will be extended as further validity studies are completed (§4.5).

### 4.2 Content validity

Content validity refers to the alignment between the items in an assessment and the construct it is designed to measure. For the CENTURY Academic Assessments, the construct for each assessment is the National Curriculum Programme of Study for the named subject.

Content validity is established through three features of the development process:

- **Blueprint-based design** (§2.2). Each assessment is built against a blueprint that maps the National Curriculum Programme of Study statements for the year group to curriculum domains, sets proportional weightings for those domains based either on the relative weight of statements within the Programme of Study or on the published Key Stage statutory test framework, and specifies the item count, target difficulty distribution and complexity profile for the assessment.
- **Item-level curriculum mapping** (§2.3). Every item is mapped to a specific National Curriculum Programme of Study statement and to the corresponding CENTURY learning nugget. The blueprint records this mapping and is used to verify that every domain is represented in the final assessment in proportion to its planned weighting.
- **Expert authoring and independent review** (§2.3, §2.4). Items are authored by experienced subject specialists from CENTURY's curriculum team and reviewed by a different specialist within the team or by the relevant subject lead. Items receive an additional check by the English team for SPaG, in addition to the standard curriculum-team review. Items are checked against the blueprint at review, including alignment to the intended Programme of Study statement and the intended complexity-profile cell.

### 4.3 Internal structure

Evidence of internal structure addresses the extent to which the items within an assessment work together to measure a single, coherent construct. The CENTURY Academic Assessments are designed as unidimensional measures of attainment within the named subject for the relevant year group. Evidence supporting this design includes:

- **IRT model fit** (§2.6.3). Item characteristics are estimated using a two-parameter logistic IRT model, which assumes a single underlying latent ability dimension. The model is fitted to all six trial forms together for each subject and year-pair, and item characteristic curves are inspected against decile-grouped raw data to provide a visual check of model fit.
- **Item discrimination** (§2.6.3, §2.6.4). Each item is required to discriminate adequately between higher-attaining and lower-attaining students before being selected for the live assessment. Items with poor discrimination (corrected total-item correlation below approximately 0.3, IRT slope  $\leq 0.4$ ) are flagged for review or discarded.
- **Marginal IRT reliability and Cronbach's alpha** (§2.6.5, §3.1). The reliability statistics reported for each assessment measure the consistency with which the items within it produce a coherent overall score. Reliability values above 0.85 (IRT marginal) and 0.80

(Cronbach's alpha) provide evidence that the items function together as a single instrument.

- **Parallel-form equivalence** (§2.6.5). The two parallel forms within each year-pair are constructed to be matched on mean IRT difficulty, expected raw score, and the spread of item difficulty. This produces two equivalent measurements of the same underlying construct.

#### 4.4 Fairness

Fairness in assessment refers to the principle that the construct should be measured equally well across different groups of students, independent of any group-level differences in mean attainment. Fairness evidence for the CENTURY Academic Assessments is provided through three components of the development process:

- **Accessibility in item authoring** (§2.3). Items are authored to be accessible to the full range of students within the target year group. Vocabulary is kept in line with reading expectations typical for the year group; items consciously avoid extraneous cultural capital requirements that would create a barrier to accessing the question; colour is checked for contrast against accessibility guidelines where it is used for a functional purpose.
- **National representativeness of the standardisation sample** (§2.7.2). Where the achieved standardisation sample deviates significantly from the national population on a dimension associated with attainment, sample weights are applied during the standardisation calculations so that the effective sample reflects the national population on that dimension. This ensures that standardised scores are referenced to a population that fairly represents students across the weighted dimensions.
- **Differential item functioning** (§3.4). Each assessment is examined for items that may function differently for different groups of students. DIF is currently analysed by sex and by eligibility for the Pupil Premium grant, with the methodology extending to additional dimensions as data quality permits. Items showing large DIF are removed; moderate DIF triggers expert review.

These components together support the use of standardised scores as fair measures of attainment across the groups examined.

#### 4.5 Ongoing validity evidence

This document records the validity evidence available at the date of publication. Further evidence, including concurrent and predictive validity studies relating CENTURY Academic

Assessment scores to other measures of student attainment, will be added as it becomes available.

## 5. Score reporting and interpretation

### 5.1 Reports in the CENTURY ecosystem

The CENTURY Academic Assessments produce reports at two levels:

- **Student-level reports**, providing each student's raw score, standardised score, stanine, topic-level breakdown, and personalised follow-up nugget recommendations.
- **Cohort-level reports**, aggregating the class or year-group's mean standardised score, mean stanine, the full stanine distribution across the cohort, topic-level performance against the national average, and recommended nuggets for the cohort.

Reports are produced automatically when an assessment window closes and are available to teachers and senior leaders through the CENTURY platform.

The reports are designed to sit within CENTURY's broader teaching and learning ecosystem, which follows a continuous four-step loop:

1. **Assess.** Establish the student's current knowledge and ability. The CENTURY Academic Assessments serve this function: standardised scores benchmark each student against the national cohort, the stanine and confidence interval describe where they sit in the distribution, and the topic-level breakdown identifies relative strengths and weaknesses against the curriculum.
2. **Diagnose.** Identify learning gaps and misconceptions precisely to inform teaching and learning. The topic-level breakdown and question-level analysis in each report show which curriculum domains and which specific items the student or cohort answered incorrectly, enabling teachers to pinpoint the underlying gaps.
3. **Empower.** Deliver personalised, curriculum-aligned content and resources to address needs. Each report links the identified gaps to specific CENTURY nuggets/micro-lessons, so that students and teachers can act on the diagnosis using the platform's wider teaching and learning resources.
4. **Consolidate.** Track progress through ongoing formative assessment and offer individualised feedback. Subsequent assessments and formative tasks on the platform allow teachers to measure whether the interventions are working, and to recalibrate as needed.

The interpretation guidance in §5.2 to §5.4 covers how to read scores at the individual, cohort and progress-over-time levels. This supports the Assess and Diagnose stages of the loop; the personalised nugget recommendations and follow-on resources support Empower and Consolidate.

## 5.2 Interpreting individual scores

For an individual student, the assessment produces:

- a **raw score**: the number of items answered correctly;
- a **standardised score**, set on a scale between 70 and 130, with a mean of 100 and standard deviation of 15, representing the student's position relative to the national standardisation cohort for the same assessment;
- a **stanine** (1 to 9), providing a coarser-grained band classification of the standardised score;
- a **topic-level breakdown**, showing the percentage of items answered correctly within each curriculum domain, compared with the national average for that domain;
- **personalised follow-up recommendations** in the form of CENTURY nuggets/micro-lessons linked to the items the student answered incorrectly.

A student's standardised score should be interpreted alongside the confidence interval associated with it (§3.3). The standardised score gives a single best estimate of attainment; the confidence interval indicates the range within which the student's underlying ability is most plausibly located. When comparing two students, or comparing a student against a benchmark, overlapping confidence intervals should temper conclusions about meaningful differences.

The topic-level breakdown should be interpreted as diagnostic information rather than as additional standardised scores. The percentage-correct figure in each topic is based on a small number of items (typically four to ten) and is therefore subject to greater measurement error than the overall standardised score. A topic showing a substantially lower percentage correct than the national average suggests a candidate area for follow-up; the recommended nuggets are designed to support this.

## 5.3 Interpreting cohort results

Cohort reports aggregate student-level scores at the class or year-group level. The headline figures are:

- the **cohort mean standardised score**, directly comparable with the national mean of 100;
- the **cohort mean stanine**, providing a quick band classification of the cohort's overall position relative to the national distribution;
- the **stanine distribution** across the cohort, showing what proportion of students fall into each of the nine bands.

A cohort with a mean standardised score close to 100 is performing, on average, at the national level. A mean above 100 indicates the cohort is performing above the national level and a mean below 100 indicates the cohort is performing below the national level. The stanine distribution shows the spread of attainment within the cohort: cohorts with similar means may still differ substantially in spread, with implications for differentiation and stretch.

Cohort topic-level breakdowns are reported as percentages correct per domain, alongside the national average for the same domain. Domains where the cohort is substantially below the national average are highlighted as areas for focus; domains substantially above are highlighted as relative strengths.

Cohort-level interpretation supports school-level planning rather than individual student intervention. Schools are best placed to interpret cohort results in the context of their own intake, prior attainment data, and curriculum delivery.

#### 5.4 Interpreting progress over time

Standardised scores referenced to the same scale (mean 100, standard deviation (SD) 15) can be compared across assessment occasions to track student progress over time. Progress can be interpreted within a year (Baseline to End-of-Year, where both are sat) and across years (End-of-Year to the following year's Baseline, or year-on-year comparisons of the same assessment occasion).

A student's standardised score represents their position relative to the national cohort sitting the same assessment at the same point in the academic year. As the scale is consistent across assessments, scores are directly comparable across occasions:

- a **stable standardised score** across consecutive assessments indicates expected progress: the student is keeping pace with the national cohort and the expected progression in the curriculum;
- a **rising standardised score** across consecutive assessments indicates faster progress than the national cohort: the student is moving up the relative distribution;

- a **falling standardised score** across consecutive assessments indicates slower progress than the national cohort: the student is falling behind relative to peers, even if their raw attainment is still increasing.

Worked examples of trajectories across three consecutive years are shown in Table 4.

**Table 4: Illustrative interpretation of standardised score trajectories**

| Student | Year 3 | Year 4 | Year 5 | Interpretation                                                                                                                                      |
|---------|--------|--------|--------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| A       | 98     | 101    | 100    | Middle-attaining student making expected progress each year compared to peers.                                                                      |
| B       | 83     | 95     | 108    | Substantial improvement over two years; well above expected progress compared to peers.                                                             |
| C       | 101    | 103    | 89     | A significant drop in Year 5; the student is falling behind the rest of the cohort and warrants investigation for contextual or attainment factors. |
| D       | 116    | 103    | 120    | Higher-attaining student with a dip in Year 4 (likely to be contextual); the Year 5 result confirms the underlying attainment level.                |

Confidence intervals (§3.3) should be considered when interpreting score changes. A difference of fewer than four SEMs (20 standardised points) between two observed scores may reflect measurement error rather than a real shift in underlying attainment.

### 5.5 Why standardised, not age-standardised

Some standardised assessments report **age-standardised scores**, in which raw scores are adjusted for the within-year variation in a student's chronological age (typically expressed in years and months) before being placed on the standardised scale. The CENTURY Academic Assessments do not adjust for age.

This is a deliberate methodological choice and reflects the purpose of the assessments. The CENTURY Academic Assessments measure attainment against the National Curriculum expectations for a year group, not generic cognitive ability or reading age. A student is compared against the national cohort of students sitting the same assessment at the same point in the academic year, against the same year-group expectations, regardless of within-year age variation.

Age-standardisation is appropriate where the underlying construct (cognitive ability, reading age) develops naturally with age and where the assessment population spans a range of ages. For curriculum-aligned attainment assessments at a specific point in a year, the relevant

comparison is to peers in the same year group at the same point in the year, which is what the standardised score provides directly.

## 5.6 Common use cases

The CENTURY Academic Assessments support a range of use cases at the student, class and cohort level. The most common are set out below:

- **End-of-Year summative reporting.** The End-of-Year assessment provides a summary measure of attainment over the academic year, suitable for inclusion in end-of-year reports, transition information for the next teacher, school-level cohort tracking, and as part of academic reporting to parents and guardians.
- **Baseline diagnostic at the start of the year.** The Baseline assessment establishes students' starting points at the beginning of the year, identifies retention gaps from the previous year, and supports planning by the new teacher.
- **Measuring *summer learning loss* (reduction of retrieval strength).** Where students sit both the End-of-Year assessment in the summer term and the corresponding Baseline assessment in the following autumn (the two are equated parallel forms, §1.2), the change in standardised score between them provides a direct measure of summer retention or attainment loss.
- **End-of-term summary.** Either assessment can be used as a single-sitting summative measure at the end of a term, suitable for tracking progress between terms across the year.
- **Casual mid-year admissions.** Students joining a school mid-year can sit the Baseline assessment for their current year group to provide the receiving teacher with a standardised reference point for the new student, alongside the rest of the cohort's existing Baseline data.
- **Between-year progress tracking.** Standardised scores referenced to the same scale across assessments support direct progress tracking from year to year, as illustrated in §5.4.

## 5.7 Limits of interpretation

The CENTURY Academic Assessments are designed to inform classroom-level decisions about teaching, intervention and reporting. When interpreting results:

- A standardised score is a single best estimate of a student's attainment on the day of the assessment, subject to measurement error (§3.3). Confidence intervals should be considered when interpreting individual scores and when comparing students.
- Topic-level breakdowns are diagnostic indicators rather than standardised scores in their own right. They are based on small numbers of items and should be used to highlight candidate areas for follow-up, not to make precise judgements about a student's relative attainment within domains.
- Standardised scores are referenced to the national cohort sitting the same assessment at the same point in the academic year. They do not measure absolute mastery against a fixed criterion.
- Results from the CENTURY Academic Assessments should be considered alongside other classroom evidence when teachers and schools draw conclusions about individual students or cohorts. The assessments are not designed as a substitute for national statutory assessments (§1.1).

## 6. Limitations and ongoing development

### 6.1 Suite development status

The CENTURY Academic Assessments suite is being built out in stages. Each assessment passes through the full development process described in §2: blueprint, item authoring and review, item trial, item analysis and selection, final form construction, and standardisation trial. Table 1 lists the 35 assessments in the suite. Table 5 sets out the stages of the assessment development process, indicating the current stage of each assessment (at the date of publication).

**Table 5: Development status at August 2026**

| Stage                                          | Assessments                                              | Evidence available now                                                                                                        | Next milestone                                                                            | Stage                                          |
|------------------------------------------------|----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|------------------------------------------------|
| Standardised                                   | Year 5 and Year 6 Baseline (6 tests)                     | Full Results Supplement: item trial, standardisation sample, reliability, SEM, confidence intervals, DIF and item performance | First scheduled re-standardisation, autumn 2028                                           | Standardised                                   |
| Standardisation complete, analysis in progress | Year 2, Year 3, Year 4 and Year 5 End-of-Year (11 tests) | Standardisation complete during summer 2026; No technical results yet                                                         | Result Supplements published in the spring 2027 update to this Manual, in time for use in | Standardisation complete, analysis in progress |

|                                                  |                                               |                                                                                                  |                                                                                                                          |                                                  |
|--------------------------------------------------|-----------------------------------------------|--------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------|
|                                                  |                                               | to be published                                                                                  | the summer 2027 term                                                                                                     |                                                  |
| In standardisation                               | Year 3, Year 4 and Year 7 Baselines (7 tests) | Item trial complete; standardisation trial in progress in the first half of the autumn 2026 term | Result Supplements published in the spring 2027 update to this Manual                                                    | In standardisation                               |
| Item trial complete with results to be processed | Year 6 End-of-Year (3 tests)                  | Item trial complete during summer 2026; awaiting processing, needed for summer 2027              | Standardisation scheduled in the term each form is intended to be used (§2.7.1); Supplements published as each completes | Item trial complete with results to be processed |
| Item trial                                       | The remaining assessments in Table 1          | Item-level statistics from the trial, held internally and used for form construction             | Standardisation scheduled in the term each form is intended to be used (§2.7.1); Supplements published as each completes | Item trial                                       |

Standardised scores and stanines are only available for an assessment once its standardisation analysis is complete and its score conversion published, because both are derived from that conversion (§2.7.3). Assessments at earlier stages report raw scores, curriculum-domain breakdowns and question-level analysis, without a standardised score or a comparison against the national average. The status of each assessment, and the dates of its most recent item trial and standardisation trial, are recorded in its Results Supplement once published.

End-of-Year and Baseline assessments are standardised in different academic terms (§2.7.1). The two forms of an equated parallel-form pair are therefore standardised separately, in line with the term in which each form is intended to be used, so the two halves of a pair may sit at different stages in Table 5.

## 6.2 Sample composition

The standardisation samples achieved to date have shown small deviations from the national population on some demographic dimensions. Where present, deviations are reported in full in the relevant Results Supplement, alongside the chi-square comparisons and any sample weights applied during the standardisation calculations (§2.7.2).

The most consistent deviations observed have been:

- **Geographical distribution.** Achieved samples have been weighted towards the North of England relative to the national distribution. Region is not strongly associated with attainment once Pupil Premium eligibility is accounted for in our standardisation

samples, so no weighting on region is applied; however, schools in under-represented regions are actively being targeted in recruitment for subsequent standardisation trials.

- **Pupil Premium eligibility.** Achieved samples have included higher proportions of students eligible for the Pupil Premium grant than the national population. Since Pupil Premium is associated with attainment, sample weights are applied during standardisation so that the effective sample reflects the national Pupil Premium proportion.
- **English as an Additional Language (EAL).** Higher proportions of EAL students have been observed in some samples. The binary EAL classification covers a wide range of language fluency, and language fluency rather than EAL status is likely to be the variable most strongly associated with attainment; this is monitored at each standardisation.

These limitations are reported transparently in each Results Supplement so that schools can take them into account when interpreting standardised scores.

### 6.3 Post-pandemic standardisation

The CENTURY Academic Assessments have been standardised on samples sat in autumn 2025 and subsequent terms, all of which post-date the educational disruption caused by the COVID-19 pandemic. This contrasts with some commercially available standardised assessments that continue to rely on pre-pandemic standardisations, the validity of which may be affected by documented changes in national student attainment.

National attainment patterns may continue to evolve in response to factors such as curriculum revision, post-pandemic learning loss and recovery, and changes in teaching practice. The rolling re-standardisation programme described in §6.4 is designed to keep CENTURY's standardised scores referenced to a contemporary national cohort.

### 6.4 Re-standardisation and ongoing review

Each assessment in the suite is reviewed as part of CENTURY's ongoing programme of review:

- **Item-level statistics** from live administrations are monitored against the original trial statistics. Items showing changes in functioning over time are flagged for review and, where necessary, replaced.
- **Cohort-level statistics** from live administrations (mean standardised scores, distributions, reliability) are monitored against the standardisation sample statistics. Significant drift may indicate that the underlying national distribution has shifted, triggering re-standardisation.

- **Re-standardisation** of each assessment is scheduled on a defined cycle to refresh the score conversions against a contemporary national cohort.

The frequency and outcomes of these reviews for each assessment are reported in the relevant Results Supplement.

## 6.5 Future development

Further work planned for the suite includes:

- **Extending the DIF analysis** (§3.4) to English as an additional language and special educational needs, as the quality of demographic data in the standardisation samples improves.
- **Concurrent validity studies** (§4.5) relating CENTURY Academic Assessment scores to other measures of student attainment, as opportunities arise.
- **Completion of the suite**, including standardisation of currently outstanding End-of-Year forms and any additional year-group or subject coverage agreed in subsequent product roadmaps.
- **Publication of Results Supplements** for more of the assessments will be in the spring 2027 update to this Manual, following completion of their standardisation trials in the summer and autumn terms of 2026.

## 7. Glossary and references

### 7.1 Glossary

**Anchor item.** An item that appears in more than one trial form, used to link item difficulty estimates across forms.

**Assessment.** A single CENTURY Academic Assessment for a specific subject and year-pair. Each is available in two equated parallel forms (Baseline and End-of-Year).

**Baseline assessment.** The form of an assessment administered at the start of an academic year to establish students' starting points (§1.2).

**Benjamini–Hochberg procedure.** A statistical method for adjusting p-values to control the false discovery rate when multiple statistical tests are conducted simultaneously. Used in CENTURY's DIF analysis (§3.4).

**Blueprint.** The foundational design document for each assessment, specifying its structure, content, balance, and trial design (§2.2).

**Cartwheel design.** Informal name for the spiralled trial design used in CENTURY's item trials (see *spiralled trial design*).

**Cohort.** A defined group of students. Used in two senses in this manual: a school cohort (a class or year-group sitting an assessment together) and the national standardisation cohort (the national sample against which standardised scores are referenced).

**Confidence interval (CI).** A range around an observed standardised score within which the student's underlying ability is likely to fall, accounting for measurement error. The 95% CI spans approximately two SEMs either side of the observed score (§3.3).

**Construct.** The skill or knowledge domain that an assessment is designed to measure. For each CENTURY Academic Assessment, the construct is the year group's National Curriculum Programme of Study for the named subject.

**Corrected total-item correlation.** Classical item discrimination index: the correlation between a student's score on a single item and their total score on the assessment excluding that item.

**Cronbach's alpha.** A classical internal-consistency reliability coefficient, ranging from 0 to 1; higher values indicate that items within an assessment provide more consistent measurement of the same construct (§3.1).

**Curriculum domain.** A high-level grouping of curriculum content within a subject (for example, *Number and Place Value* in Mathematics or *States of Matter* in Science).

**Differential Item Functioning (DIF).** A property whereby an item has a higher or lower probability of being answered correctly by students from one group, compared with students of the same ability in another group (§3.4).

**End-of-Year assessment.** The form of an assessment administered at the end of an academic year to summarise attainment over the year (§1.2).

**Equated parallel forms.** Two forms of an assessment built from a common item pool, calibrated to the same difficulty and balanced across the same curriculum domains, but differing in the specific items they contain (§1.2).

**Exact-answer item.** An item type in which the student types their answer into a response box, marked by a Numeric, Phonetic, or Exact algorithm, depending on the construct being assessed (§2.3).

**Facility (facility value).** The percentage of students who answered a given item correctly. A classical item statistic indicating item difficulty.

**Item Characteristic Curve (ICC).** A plot of the probability of correctly answering an item as a function of underlying ability ( $\theta$ ), derived from the IRT model.

**Item Response Theory (IRT).** A family of statistical models that describe how the probability of a correct response to an item depends on a student's underlying ability and the characteristics of the item.

**IRT difficulty (b parameter).** The location parameter in a 2-PL IRT model: the value of  $\theta$  at which a student has a 50% probability of correctly answering the item.

**IRT slope (a parameter).** The discrimination parameter in a 2-PL IRT model: the slope of the item characteristic curve at the inflection point. Higher values indicate sharper discrimination between adjacent ability levels.

**Labelling item.** An item type in which the student attaches a set of provided labels to specified points on a diagram or image (§2.3).

**Marginal IRT reliability.** A reliability statistic derived from an IRT model, summarising the consistency of the test information across the full ability range. Calculated as part of the final form construction check (§2.6.5).

**Matching item.** An item type in which the student pairs items from two columns.

**Multiple-choice item.** An item type in which the student selects a single correct answer from a fixed set of options (§2.3).

**National Curriculum Programme of Study.** The statutory document published by the Department for Education setting out what students in each year group of state-funded schools in England should be taught in each subject.

**Not-reached rate.** The percentage of students who did not reach a given item before time expired.

**Nugget.** A micro-lesson of curriculum-aligned learning content on the CENTURY platform.

**Omit-of-seen rate.** The percentage of students who reached a given item but chose not to attempt it.

**Parallel forms.** See *Equated parallel forms*.

**Raw score.** The number of items answered correctly on an assessment.

**Reliability.** The consistency of scores produced by an assessment across replications. Reliability for the CENTURY Academic Assessments is reported as both IRT marginal reliability and Cronbach's alpha (§3.1).

**Spiralled trial design.** A trial design in which items rotate across multiple overlapping forms, so that every item is administered to a subset of the trial sample and item characteristics can be estimated relative to all other items. Informally referred to as a cartwheel design (§2.6.1).

**Standard Error of Measurement (SEM).** A statistic quantifying the typical magnitude of measurement error around a single observed score. Calculated as  $SD \times \sqrt{1 - \alpha}$ , where SD is the standard deviation of the standardised score scale and  $\alpha$  is the reliability (§3.2).

**Standardisation.** The process of converting raw scores to standardised scores against a normative sample. CENTURY uses a model fitted with Taylor polynomials (§2.7.3).

**Standardised score.** A score on a scale with a mean of 100 and a standard deviation of 15, representing a student's position relative to the national standardisation cohort for the same assessment.

**Stanine.** A coarser-grained band classification of standardised scores on a 1-to-9 scale, derived from cut-points at half-standard-deviation intervals (§2.7.4).

**Test information function.** A function of underlying ability ( $\theta$ ), showing the precision with which the assessment measures at each point on the ability range. Used in final form construction to ensure information is spread across the relevant ability range (§2.6.5).

**Two-parameter logistic (2-PL) model.** The IRT model used in CENTURY's item analysis and DIF analysis, in which each item is described by a discrimination (a) and difficulty (b) parameter (§2.6.3, §3.4).

**Validity.** The degree to which evidence and theory support the interpretations of standardised scores from an assessment for its intended uses (§4).

**Year-pair.** A pair consisting of a Year N End-of-Year assessment and the corresponding Year N+1 Baseline assessment. The two are equated parallel forms (§1.2).

## 7.2 References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>

Department for Education. *Explore education statistics*.  
<https://explore-education-statistics.service.gov.uk/>

Department for Education. *Get information about schools*.  
<https://get-information-schools.service.gov.uk/Downloads>

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, 14(4), 329–349. [https://doi.org/10.1207/S15324818AME1404\\_2](https://doi.org/10.1207/S15324818AME1404_2)

Lenhard, A., Lenhard, W., & Gary, S. (2018). *Continuous Norming (cNORM)*. The Comprehensive R Network. <https://CRAN.R-project.org/package=cNORM>

Willse, J. T. (2018). *CTT: Classical Test Theory Functions*. R package version 2.3.3.  
<https://CRAN.R-project.org/package=CTT>

## Assessment Supplements

A Results Supplement is published for each assessment once it has completed standardisation. Each reports that assessment's item trial, standardisation sample and composition against the national population, reliability, standard error of measurement, differential item functioning and item performance, and refers to this Manual for the methodology.

Supplements for the standardised assessments follow. Supplements for the Year 3, Year 4 and Year 7 Baseline assessments and the Year 2 to Year 5 End-of-Year assessments will be added in the spring 2027 update, and for the remaining assessments as each completes standardisation (§6.1).

### Year 5 Baseline SPaG

This supplement reports the technical results for the Year 5 Baseline SPaG assessment. It should be read alongside the CENTURY Academic Assessments Technical Manual, which sets out the methodology used to develop, trial, standardise and report all assessments in the suite. Section references in this document refer to that Manual.

#### Assessment overview

The Year 5 Baseline SPaG assessment measures students' attainment at the start of Year 5 against the Year 4 Spelling, Punctuation and Grammar (SPaG) Programme of Study of the English National Curriculum.

The assessment comprises **30 items**, designed to be completed within a single 50-minute sitting and covers the national curriculum specification for grammar, punctuation and vocabulary.

The Year 5 Baseline SPaG assessment is the autumn-term parallel form of the Year 4 End-of-Year SPaG content set. The two forms are equated parallel forms (Manual §1.2): items in each were drawn from a common trialled pool, calibrated to the same difficulty and balanced across the same curriculum domains. The Year 4 End-of-Year SPaG assessment was standardised separately during the summer term.

#### Item trial

In the summer term of 2025, **120 items** were trialled with **651 Year 4 students** (pre-standardisation testing, which is detailed below) across six spiralled forms (Manual §2.6.1). The trialled item pool fed selection of items into both the Year 4 End-of-Year SPaG and Year 5 Baseline SPaG assessments. Classical item statistics and Item Response Theory (IRT) parameters were calculated for all trialled items, with all six trial forms analysed together (Manual §2.6.3). Of these, **30 items** were selected for the Year 5 Baseline SPaG assessment, balanced for curriculum domain coverage, mean IRT difficulty and spread of difficulty (Manual

§2.6.5). The selected items achieved a marginal IRT reliability of **0.891** at final form construction.

### Standardisation

The assessment was standardised on **1,598 students** across 44 schools during September and October 2025 (the start of the academic year in England). Raw scores ranged from 0 to 30, with a mean of 15.3 (standard deviation 6.9) and a median of 15. The interquartile range (25th to 75th percentile) was 10 to 20.

The achieved sample composition was compared with national population figures published by the Department for Education using one-sample chi-square tests. Results are shown in Table 1.<sup>1</sup>

**Table 1: Sample composition vs national population**

| Dimension                     | Category | Population N | Population % | Sample N | Sample % | p-value |
|-------------------------------|----------|--------------|--------------|----------|----------|---------|
| Sex: Year 5                   | Male     | 317,735      | 50.8%        | 774      | 49.6%    | 0.317   |
|                               | Female   | 307,128      | 49.2%        | 787      | 50.4%    |         |
|                               | Unknown  | 0            | —            | 37       | —        |         |
| Pupil Premium: Primary sector | PP       | 1,126,353    | 24.7%        | 496      | 31.9%    | < 0.001 |
|                               | Non-PP   | 3,428,621    | 75.3%        | 1,060    | 68.1%    |         |
|                               | Unknown  | 0            | —            | 42       | —        |         |
| EAL: Primary sector           | EAL      | 1,065,551    | 23.4%        | 408      | 25.5%    | < 0.001 |
|                               | Non-EAL  | 3,474,334    | 76.3%        | 1,137    | 71.2%    |         |
|                               | Unknown  | 15,089       | 0.3%         | 53       | 3.3%     |         |
| SEN: Year 5                   | SEN      | 132,027      | 21.2%        | 311      | 21.3%    | 0.872   |
|                               | Non-SEN  | 492,836      | 78.9%        | 1,149    | 78.7%    |         |
|                               | Unknown  | 0            | —            | 138      | —        |         |
| Region: Year 5                | North    | 180,214      | 28.8%        | 837      | 52.4%    | < 0.001 |

|  |          |         |        |     |        |  |
|--|----------|---------|--------|-----|--------|--|
|  | Midlands | 194,965 | 31.20% | 412 | 25.80% |  |
|  | South    | 249,684 | 40.00% | 349 | 21.80% |  |

- Sex and SEN were not significantly different from the national population ( $p = 0.317$  and  $p = 0.872$  respectively). The sample is representative on these dimensions.
- Pupil Premium eligibility was over-represented in the sample (31.9% vs 24.7% nationally,  $p < 0.001$ ).
- English as an Additional Language was over-represented in the sample (25.5% vs 23.4% nationally,  $p < 0.001$ ).
- The sample was geographically skewed towards the North of England (52.4% vs 28.8% nationally). Region is not strongly associated with attainment after other demographic factors are taken into account, and no weighting on region is applied (Manual §6.2).

Sample weights were applied during standardisation so that the effective sample matched the national population proportion for Pupil Premium eligibility (Manual §2.7.2). No other weights were applied.

### Reliability and SEM

The internal-consistency reliability of the assessment, measured by Cronbach's alpha, is **0.896** – above the target of 0.80 (Manual §3.1). This indicates that the items within the assessment work together coherently as a measure of the same underlying construct, with high consistency across the standardisation sample.

The Standard Error of Measurement is **4.8 standardised score points**. For a student with an observed standardised score of 100, the 95% confidence interval around their true score is approximately 90.6 to 109.4, and the 68% confidence interval is 95.2 to 104.8 (Manual §3.3). If the student were to sit comparable assessments many times, around 95 in every 100 of those scores would fall within the 95% range.

### Differential Item Functioning

Items in the Year 5 Baseline SPaG assessment were examined for differential item functioning (DIF) by sex and by Pupil Premium eligibility, using a 2-PL multiple-group IRT model with Benjamini–Hochberg adjustment for multiple testing (Manual §3.4). The number of items showing statistically significant DIF is shown in Table 2.

**Table 2: Items showing significant DIF**

| Dimension     | Items with significant DIF | Total items | Proportion |
|---------------|----------------------------|-------------|------------|
| Sex           | 3                          | 30          | 10.00%     |
| Pupil Premium | 0                          | 30          | 0.00%      |

The magnitude of DIF for each flagged item was classified using the scale of Jodoin and Gierl (2001)<sup>8</sup> as either negligible or moderate. The items showing DIF were reviewed and no obvious patterns of bias were found; the flagged items spanned different curriculum domains, skills and item types.

#### Item performance summary

All 30 items in the Year 5 Baseline SPaG assessment passed the item-selection discard rules at the item-trial stage (Manual §2.6.4): each item had a facility within 10–90%, an omit-of-seen rate below 15%, an IRT slope above 0.4, and an IRT difficulty parameter within the range -3 to +3. The spread of facility values across the assessment supports measurement of attainment across the full ability range targeted by the Year 5 Baseline SPaG Assessment (Manual §2.6.5).

The following summary statistics describe item performance in the standardisation sample (n = 1,598).

**Table 3: Item-level summary statistics**

| Item-level statistic             | Mean   | Range         |
|----------------------------------|--------|---------------|
| Facility (% correct)             | 51.10% | 18.7% – 89.5% |
| Corrected total-item correlation | 0.45   | 0.21 – 0.65   |
| Omission rate                    | 10.90% | 1.4% – 19.5%  |
| Not-reached rate                 | 6.20%  | 0.0% – 10.7%  |

Items showing any concerning indicators in the standardisation sample (for example, a corrected total-item correlation below 0.3 or an omit-of-seen rate above 15%) are flagged for review at the next re-standardisation cycle (Manual §6.4).

<sup>8</sup> Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, 14(4), 329–349. [https://doi.org/10.1207/S15324818AME1404\\_2](https://doi.org/10.1207/S15324818AME1404_2)

## Year 5 Baseline Mathematics

This supplement reports the technical results for the Year 5 Baseline Mathematics assessment. It should be read alongside the CENTURY Academic Assessments Technical Manual, which sets out the methodology used to develop, trial, standardise and report all assessments in the suite. Section references in this document refer to that Manual.

### Assessment overview

The Year 5 Baseline Mathematics assessment measures students' attainment at the start of Year 5 against the Year 4 Mathematics Programme of Study of the English National Curriculum.

The assessment comprises **28 items**, designed to be completed within a single 50-minute sitting and covers the following topics: Addition and Subtraction, Fractions (including decimals), Measurement, Multiplication and Division, Number and Place Value, Position and Direction, Properties of Shapes and Statistics.

The Year 5 Baseline Mathematics assessment is the autumn-term parallel form of the Year 4 End-of-Year Mathematics content set. The two forms are equated parallel forms (Manual §1.2): items in each were drawn from a common trialled pool, calibrated to the same difficulty and balanced across the same curriculum domains. The Year 4 End-of-Year Mathematics assessment was standardised separately during the summer term.

### Item trial

In the summer term of 2025, **84 items** were trialled with **818 Year 4 students** (pre-standardisation testing, which is detailed below) across six spiralled forms (Manual §2.6.1). The trialled item pool fed selection of items into both the Year 4 End-of-Year and Year 5 Baseline Mathematics assessments. Classical item statistics and Item Response Theory (IRT) parameters were calculated for all trialled items, with all six trial forms analysed together (Manual §2.6.3). Of these, **28 items** were selected for the Year 5 Baseline Mathematics assessment, balanced for curriculum domain coverage, mean IRT difficulty and spread of difficulty (Manual §2.6.5). The selected items achieved a marginal IRT reliability of **0.891** at final form construction.

### Standardisation

The assessment was standardised on **1,917 students** across 52 schools during September and October 2025 (the start of the academic year in England). Raw scores ranged from 0 to 28, with a mean of 13.3 (standard deviation 6.3) and a median of 13. The interquartile range (25th to 75th percentile) was 8 to 18.

The achieved sample composition was compared with national population figures published by the Department for Education using one-sample chi-square tests. Results are shown in Table 1.

**Table 1: Sample composition vs national population**

| Dimension                     | Category | Population N | Population % | Sample N | Sample % |
|-------------------------------|----------|--------------|--------------|----------|----------|
| Sex: Year 5                   | Male     | 317,735      | 50.80%       | 935      | 49.70%   |
|                               | Female   | 307,128      | 49.20%       | 946      | 50.30%   |
|                               | Unknown  | 0            | —            | 36       | —        |
| Pupil Premium: Primary sector | PP       | 1,126,353    | 24.70%       | 622      | 33.20%   |
|                               | Non-PP   | 3,428,621    | 75.30%       | 1,254    | 66.80%   |
|                               | Unknown  | 0            | —            | 41       | —        |
| EAL: Primary sector           | EAL      | 1,065,551    | 23.40%       | 544      | 28.40%   |
|                               | Non-EAL  | 3,474,334    | 76.30%       | 1,320    | 68.90%   |
|                               | Unknown  | 15,089       | 0.30%        | 53       | 2.80%    |
| SEN: Year 5                   | SEN      | 132,027      | 21.20%       | 372      | 21.10%   |
|                               | Non-SEN  | 492,836      | 78.90%       | 1,394    | 78.90%   |
|                               | Unknown  | 0            | —            | 151      | —        |
| Region: Year 5                | North    | 180,214      | 28.80%       | 1,024    | 53.40%   |
|                               | Midlands | 194,965      | 31.20%       | 528      | 27.50%   |
|                               | South    | 249,684      | 40.00%       | 365      | 19.00%   |

- Sex and SEN were not significantly different from the national population ( $p = 0.322$  and  $p = 0.947$  respectively). The sample is representative on these dimensions.
- Pupil Premium eligibility was over-represented in the sample (33.2% vs 24.7% nationally,  $p < 0.001$ ).
- English as an Additional Language was over-represented in the sample (28.4% vs 23.4% nationally,  $p < 0.001$ ).
- The sample was geographically skewed towards the North of England (53.4% vs 28.8% nationally). Region is not strongly associated with attainment after other demographic factors are taken into account, and no weighting on region is applied (Manual §6.2).

Sample weights were applied during standardisation so that the effective sample matched the national population proportion for Pupil Premium eligibility (Manual §2.7.2). No other weights were applied.

## Reliability and SEM

The internal-consistency reliability of the assessment, measured by Cronbach's alpha, is **0.894** – above the target of 0.80 (Manual §3.1). This indicates that the items within the assessment work together coherently as a measure of the same underlying construct, with high consistency across the standardisation sample.

The Standard Error of Measurement is **4.9 standardised score points**. For a student with an observed standardised score of 100, the 95% confidence interval around their true score is approximately 90.4 to 109.6, and the 68% confidence interval is 95.1 to 104.9 (Manual §3.3). If the student were to sit comparable assessments many times, around 95 in every 100 of those scores would fall within the 95% range.

## Differential Item Functioning

Items in the Year 5 Baseline Mathematics assessment were examined for differential item functioning (DIF) by sex and by Pupil Premium eligibility, using a 2-PL multiple-group IRT model with Benjamini–Hochberg adjustment for multiple testing (Manual §3.4). The number of items showing statistically significant DIF is shown in Table 2.

**Table 2: Items showing significant DIF**

| Dimension     | Items with significant DIF | Total items | Proportion |
|---------------|----------------------------|-------------|------------|
| Sex           | 8                          | 28          | 28.60%     |
| Pupil Premium | 4                          | 28          | 14.30%     |

The magnitude of DIF for each flagged item was classified using the scale of Jodoin and Gierl (2001)<sup>9</sup> as either negligible or moderate. The items showing DIF were reviewed and no obvious patterns of bias were found; the flagged items spanned different curriculum domains, skills and item types.

## Item performance summary

All 28 items in the Year 5 Baseline Mathematics assessment passed the item-selection discard rules at the item-trial stage (Manual §2.6.4): each item had a facility within 10–90%, an omit-of-seen rate below 15%, an IRT slope above 0.4, and an IRT difficulty parameter within the range -3 to +3. The spread of facility values across the assessment supports measurement of attainment across the full ability range targeted by the Year 5 Baseline Mathematics Assessment (Manual §2.6.5).

<sup>9</sup> Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, 14(4), 329–349. [https://doi.org/10.1207/S15324818AME1404\\_2](https://doi.org/10.1207/S15324818AME1404_2)

The following summary statistics describe item performance in the standardisation sample (n = 1,917).

**Table 3: Item-level summary statistics**

| Item-level statistic             | Mean   | Range         |
|----------------------------------|--------|---------------|
| Facility (% correct)             | 47.40% | 12.0% – 84.8% |
| Corrected total-item correlation | 0.45   | 0.25 – 0.61   |
| Omission rate                    | 8.10%  | 1.0% – 18.9%  |
| Not-reached rate                 | 3.50%  | 0.0% – 8.2%   |

Items showing any concerning indicators in the standardisation sample (for example, a corrected total-item correlation below 0.3 or an omit-of-seen rate above 15%) are flagged for review at the next re-standardisation cycle (Manual §6.4).

## Year 5 Baseline Science

This supplement reports the technical results for the Year 5 Baseline Science assessment. It should be read alongside the CENTURY Academic Assessments Technical Manual, which sets out the methodology used to develop, trial, standardise and report all assessments in the suite. Section references in this document refer to that Manual.

### Assessment overview

The Year 5 Baseline Science assessment measures students' attainment at the start of Year 5 against the Year 4 Science Programme of Study of the English National Curriculum.

The assessment comprises **30 items**, designed to be completed within a single 50-minute sitting and covers the following topics: Animals including humans, Electricity, Living things and their habitats, Sound, and States of matter.

The Year 5 Baseline Science assessment is the autumn-term parallel form of the Year 4 End-of-Year Science content set. The two forms are equated parallel forms (Manual §1.2): items in each were drawn from a common trialled pool, calibrated to the same difficulty and balanced across the same curriculum domains. The Year 4 End-of-Year Science assessment was standardised separately during the summer term.

### Item trial

In the summer term of 2025, **96 items** were trialled with **732 Year 4 students** (pre-standardisation testing, which is detailed below) across six spiralled forms (Manual §2.6.1). The trialled item pool fed selection of items into both the Year 4 End-of-Year Science and Year 5 Baseline Science assessments. Classical item statistics and Item Response Theory (IRT) parameters were calculated for all trialled items, with all six trial forms analysed together (Manual §2.6.3). Of these, **30 items** were selected for the Year 5 Baseline Science assessment, balanced for curriculum domain coverage, mean IRT difficulty and spread of difficulty (Manual §2.6.5). The selected items achieved a marginal IRT reliability of **0.864** at final form construction.

### Standardisation

The assessment was standardised on **1,396 students** across 38 schools during September and October 2025 (the start of the academic year in England). Raw scores ranged from 0 to 30, with a mean of 15.7 (standard deviation 5.8) and a median of 16. The interquartile range (25th to 75th percentile) was 12 to 20.

The achieved sample composition was compared with national population figures published by the Department for Education using one-sample chi-square tests. Results are shown in Table 1.

**Table 1: Sample composition vs national population**

| Dimension                     | Category | Population N | Population % | Sample N | Sample % | p-value |
|-------------------------------|----------|--------------|--------------|----------|----------|---------|
| Sex: Year 5                   | Male     | 317,735      | 50.80%       | 679      | 48.90%   | 0.143   |
|                               | Female   | 307,128      | 49.20%       | 710      | 51.10%   |         |
|                               | Unknown  | 0            | —            | 7        | —        |         |
| Pupil Premium: Primary sector | PP       | 1,126,353    | 24.70%       | 453      | 32.70%   | < 0.001 |
|                               | Non-PP   | 3,428,621    | 75.30%       | 932      | 67.30%   |         |
|                               | Unknown  | 0            | —            | 11       | —        |         |
| EAL: Primary sector           | EAL      | 1,065,551    | 23.40%       | 318      | 22.80%   | 0.068   |
|                               | Non-EAL  | 3,474,334    | 76.30%       | 1,068    | 76.50%   |         |
|                               | Unknown  | 15,089       | 0.30%        | 10       | 0.70%    |         |
| SEN: Year 5                   | SEN      | 132,027      | 21.20%       | 257      | 20.10%   | 0.364   |
|                               | Non-SEN  | 492,836      | 78.90%       | 1,022    | 79.90%   |         |
|                               | Unknown  | 0            | —            | 117      | —        |         |
| Region: Year 5                | North    | 180,214      | 28.80%       | 771      | 55.20%   | < 0.001 |
|                               | Midlands | 194,965      | 31.20%       | 405      | 29.00%   |         |
|                               | South    | 249,684      | 40.00%       | 220      | 15.80%   |         |

- Sex and SEN were not significantly different from the national population ( $p = 0.143$  and  $p = 0.364$  respectively). The sample is representative on these dimensions.
- Pupil Premium eligibility was over-represented in the sample (32.7% vs 24.7% nationally,  $p < 0.001$ ).
- English as an Additional Language differed slightly but not significantly different from the national population (22.8% vs 23.4%,  $p = 0.068$ ).
- The sample was geographically skewed towards the North of England (55.2% vs 28.8% nationally). Region is not strongly associated with attainment after other demographic factors are taken into account, and no weighting on region is applied (Manual §6.2).

Sample weights were applied during standardisation so that the effective sample matched the national population proportion for Pupil Premium eligibility (Manual §2.7.2). No other weights were applied.

### Reliability and SEM

The internal-consistency reliability of the assessment, measured by Cronbach's alpha, is **0.850** – above the target of 0.80 (Manual §3.1). This indicates that the items within the assessment work together coherently as a measure of the same underlying construct, with high consistency across the standardisation sample.

The Standard Error of Measurement is **5.8 standardised score points**. For a student with an observed standardised score of 100, the 95% confidence interval around their true score is approximately 88.6 to 111.4, and the 68% confidence interval is 94.2 to 105.8 (Manual §3.3). If the student were to sit comparable assessments many times, around 95 in every 100 of those scores would fall within the 95% range.

### Differential Item Functioning

Items in the Year 5 Baseline Science assessment were examined for differential item functioning (DIF) by sex and by Pupil Premium eligibility, using a 2-PL multiple-group IRT model with Benjamini–Hochberg adjustment for multiple testing (Manual §3.4). The number of items showing statistically significant DIF is shown in Table 2.

**Table 2: Items showing significant DIF**

| Dimension     | Items with significant DIF | Total items | Proportion |
|---------------|----------------------------|-------------|------------|
| Sex           | 3                          | 30          | 10.00%     |
| Pupil Premium | 0                          | 30          | 0.00%      |

The magnitude of DIF for each flagged item was classified using the scale of Jodoin and Gierl (2001)<sup>10</sup> as negligible. The items showing DIF were reviewed and no obvious patterns of bias were found; the flagged items spanned different curriculum domains, skills and item types.

### Item performance summary

All 30 items in the Year 5 Baseline Science assessment passed the item-selection discard rules at the item-trial stage (Manual §2.6.4): each item had a facility within 10–90%, an omit-of-seen rate below 15%, an IRT slope above 0.4, and an IRT difficulty parameter within the range -3 to +3. The spread of facility values across the assessment supports measurement of attainment

<sup>10</sup> Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, 14(4), 329–349. [https://doi.org/10.1207/S15324818AME1404\\_2](https://doi.org/10.1207/S15324818AME1404_2)

across the full ability range targeted by the Year 5 Baseline Science Assessment (Manual §2.6.5).

The following summary statistics describe item performance in the standardisation sample (n = 1,396).

**Table 3: Item-level summary statistics**

| Item-level statistic             | Mean   | Range        |
|----------------------------------|--------|--------------|
| Facility (% correct)             | 52.20% | 6.2% – 92.9% |
| Corrected total-item correlation | 0      | 0.12 – 0.55  |
| Omission rate                    | 7.20%  | 0.6% – 19.6% |
| Not-reached rate                 | 3.50%  | 0.0% – 6.7%  |

Items showing any concerning indicators in the standardisation sample (for example, a corrected total-item correlation below 0.3 or an omit-of-seen rate above 15%) are flagged for review at the next re-standardisation cycle (Manual §6.4).

## Year 6 Baseline SPaG

This supplement reports the technical results for the Year 6 Baseline SPaG assessment. It should be read alongside the CENTURY Academic Assessments Technical Manual, which sets out the methodology used to develop, trial, standardise and report all assessments in the suite. Section references in this document refer to that Manual.

### Assessment overview

The Year 6 Baseline SPaG assessment measures students' attainment at the start of Year 6 against the Year 5 Spelling, Punctuation and Grammar (SPaG) Programme of Study of the English National Curriculum.

The assessment comprises **40 items**, designed to be completed within a single 50-minute sitting and covers the national curriculum specification for grammar, punctuation and vocabulary.

The Year 6 Baseline SPaG assessment is the autumn-term parallel form of the Year 5 End-of-Year SPaG content set. The two forms are equated parallel forms (Manual §1.2): items in each were drawn from a common trialled pool, calibrated to the same difficulty and balanced across the same curriculum domains. The Year 5 End-of-Year SPaG assessment was standardised separately during the summer term.

### Item trial

In the summer term of 2025, **120 items** were trialled with **718 Year 5 students** (pre-standardisation testing, which is detailed below) across six spiralled forms (Manual §2.6.1). The trialled item pool fed selection of items into both the Year 5 End-of-Year SPaG and Year 6 Baseline SPaG assessments. Classical item statistics and Item Response Theory (IRT) parameters were calculated for all trialled items, with all six trial forms analysed together (Manual §2.6.3). Of these, **40 items** were selected for the Year 6 Baseline SPaG assessment, balanced for curriculum domain coverage, mean IRT difficulty and spread of difficulty (Manual §2.6.5). The selected items achieved a marginal IRT reliability of **0.919** at final form construction.

### Standardisation

The assessment was standardised on **1,619 students** across 55 schools during September and October 2025 (the start of the academic year in England). Raw scores ranged from 0 to 40, with a mean of 20.3 (standard deviation 10.1) and a median of 20. The interquartile range (25th to 75th percentile) was 12 to 29.

The achieved sample composition was compared with national population figures published by the Department for Education using one-sample chi-square tests. Results are shown in Table 1.

**Table 1: Sample composition vs national population**

| Dimension                     | Category | Population N | Population % | Sample N | Sample % | p-value |
|-------------------------------|----------|--------------|--------------|----------|----------|---------|
| Sex: Year 5                   | Male     | 316,321      | 50.7%        | 820      | 51.7%    | 0.459   |
|                               | Female   | 307,098      | 49.3%        | 767      | 48.3%    |         |
|                               | Unknown  | 0            | —            | 32       | —        |         |
| Pupil Premium: Primary sector | PP       | 1,126,353    | 24.7%        | 521      | 32.3%    | < 0.001 |
|                               | Non-PP   | 3,428,621    | 75.3%        | 1,090    | 67.7%    |         |
|                               | Unknown  | 0            | —            | 8        | —        |         |
| EAL: Primary sector           | EAL      | 1,065,551    | 23.4%        | 442      | 27.3%    | < 0.001 |
|                               | Non-EAL  | 3,474,334    | 76.3%        | 1,135    | 70.1%    |         |
|                               | Unknown  | 15,089       | 0.3%         | 42       | 2.6%     |         |
| SEN: Year 5                   | SEN      | 134,140      | 21.5%        | 308      | 22.0%    | 0.629   |
|                               | Non-SEN  | 489,279      | 78.5%        | 1,089    | 78.0%    |         |
|                               | Unknown  | 0            | —            | 222      | —        |         |
| Region: Year 5                | North    | 179,714      | 28.8%        | 809      | 50.0%    | < 0.001 |
|                               | Midlands | 194,777      | 31.2%        | 540      | 33.4%    |         |
|                               | South    | 248,928      | 39.9%        | 270      | 16.7%    |         |

- Sex and SEN were not significantly different from the national population ( $p = 0.459$  and  $p = 0.629$  respectively). The sample is representative on these dimensions.
- Pupil Premium eligibility was over-represented in the sample (32.3% vs 24.7% nationally,  $p < 0.001$ ).
- English as an Additional Language was over-represented in the sample (27.3% vs 23.4% nationally,  $p < 0.001$ ).
- The sample was geographically skewed towards the North of England (50.0% vs 28.8% nationally). Region is not strongly associated with attainment after other demographic factors are taken into account, and no weighting on region is applied (Manual §6.2).

Sample weights were applied during standardisation so that the effective sample matched the national population proportion for Pupil Premium eligibility (Manual §2.7.2). No other weights were applied.

## Reliability and SEM

The internal-consistency reliability of the assessment, measured by Cronbach's alpha, is **0.936** – above the target of 0.80 (Manual §3.1). This indicates that the items within the assessment work together coherently as a measure of the same underlying construct, with high consistency across the standardisation sample.

The Standard Error of Measurement is **3.8 standardised score points**. For a student with an observed standardised score of 100, the 95% confidence interval around their true score is approximately 92.6 to 107.4, and the 68% confidence interval is 96.2 to 103.8 (Manual §3.3). If the student were to sit comparable assessments many times, around 95 in every 100 of those scores would fall within the 95% range.

## Differential Item Functioning

Items in the Year 6 Baseline SPaG assessment were examined for differential item functioning (DIF) by sex and by Pupil Premium eligibility, using a 2-PL multiple-group IRT model with Benjamini–Hochberg adjustment for multiple testing (Manual §3.4). The number of items showing statistically significant DIF is shown in Table 2.

**Table 2: Items showing significant DIF**

| Dimension     | Items with significant DIF | Total items | Proportion |
|---------------|----------------------------|-------------|------------|
| Sex           | 5                          | 40          | 12.5%      |
| Pupil Premium | 0                          | 40          | 0.0%       |

The magnitude of DIF for each flagged item was classified using the scale of Jodoin and Gierl (2001)<sup>11</sup> as either negligible or moderate. The items showing DIF were reviewed and no obvious patterns of bias were found; the flagged items spanned different curriculum domains, skills and item types.

## Item performance summary

All 40 items in the Year 6 Baseline SPaG assessment passed the item-selection discard rules at the item-trial stage (Manual §2.6.4): each item had a facility within 10–90%, an omit-of-seen rate below 15%, an IRT slope above 0.4, and an IRT difficulty parameter within the range -3 to +3. The spread of facility values across the assessment supports measurement of attainment across the full ability range targeted by the Year 6 Baseline SPaG Assessment (Manual §2.6.5).

<sup>11</sup> Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, 14(4), 329–349. [https://doi.org/10.1207/S15324818AME1404\\_2](https://doi.org/10.1207/S15324818AME1404_2)

The following summary statistics describe item performance in the standardisation sample (n = 1,619).

**Table 3: Item-level summary statistics**

| Item-level statistic             | Mean  | Range         | Item-level statistic             |
|----------------------------------|-------|---------------|----------------------------------|
| Facility (% correct)             | 50.7% | 16.2% – 91.8% | Facility (% correct)             |
| Corrected total-item correlation | 1     | 0.26 – 0.66   | Corrected total-item correlation |
| Omission rate                    | 10.7% | 1.1% – 24.6%  | Omission rate                    |
| Not-reached rate                 | 6.5%  | 0.0% – 11.5%  | Not-reached rate                 |

Items showing any concerning indicators in the standardisation sample (for example, a corrected total-item correlation below 0.3 or an omit-of-seen rate above 15%) are flagged for review at the next re-standardisation cycle (Manual §6.4).

## Year 6 Baseline Mathematics

This supplement reports the technical results for the Year 6 Baseline Mathematics assessment. It should be read alongside the CENTURY Academic Assessments Technical Manual, which sets out the methodology used to develop, trial, standardise and report all assessments in the suite. Section references in this document refer to that Manual.

### Assessment overview

The Year 6 Baseline Mathematics assessment measures students' attainment at the start of Year 6 against the Year 5 Mathematics Programme of Study of the English National Curriculum.

The assessment comprises **28 items**, designed to be completed within a single 50-minute sitting and covers the following topics: Addition and Subtraction, Fractions (including decimals and percentages), Geometry (position and direction), Geometry (properties of shapes), Measurement, Multiplication and Division, Number and Place Value, and Statistics.

The Year 6 Baseline Mathematics assessment is the autumn-term parallel form of the Year 5 End-of-Year Mathematics content set. The two forms are equated parallel forms (Manual §1.2): items in each were drawn from a common trialled pool, calibrated to the same difficulty and balanced across the same curriculum domains. The Year 5 End-of-Year Mathematics assessment was standardised separately during the summer term.

### Item trial

In the summer term of 2025, **84 items** were trialled with **709 Year 5 students** (pre-standardisation testing, which is detailed below) across six spiralled forms (Manual §2.6.1). The trialled item pool fed selection of items into both the Year 5 End-of-Year Mathematics and Year 6 Baseline Mathematics assessments. Classical item statistics and Item Response Theory (IRT) parameters were calculated for all trialled items, with all six trial forms analysed together (Manual §2.6.3). Of these, **28 items** were selected for the Year 6 Baseline Mathematics assessment, balanced for curriculum domain coverage, mean IRT difficulty and spread of difficulty (Manual §2.6.5). The selected items achieved a marginal IRT reliability of **0.901** at final form construction.

### Standardisation

The assessment was standardised on **1,765 students** across 58 schools during September and October 2025 (the start of the academic year in England). Raw scores ranged from 0 to 28, with a mean of 12.9 (standard deviation 7.0) and a median of 12. The interquartile range (25th to 75th percentile) was 7 to 18.

The achieved sample composition was compared with national population figures published by the Department for Education using one-sample chi-square tests. Results are shown in Table 1.<sup>1</sup>

**Table 1: Sample composition vs national population**

| Dimension                     | Category | Population N | Population % | Sample N | Sample % | p-value |
|-------------------------------|----------|--------------|--------------|----------|----------|---------|
| Sex: Year 5                   | Male     | 316,321      | 50.7%        | 905      | 52.3%    | 0.183   |
|                               | Female   | 307,098      | 49.3%        | 824      | 47.7%    |         |
|                               | Unknown  | 0            | —            | 36       | —        |         |
| Pupil Premium: Primary sector | PP       | 1,126,353    | 24.7%        | 551      | 31.4%    | < 0.001 |
|                               | Non-PP   | 3,428,621    | 75.3%        | 1,203    | 68.6%    |         |
|                               | Unknown  | 0            | —            | 11       | —        |         |
| EAL: Primary sector           | EAL      | 1,065,551    | 23.4%        | 497      | 28.2%    | < 0.001 |
|                               | Non-EAL  | 3,474,334    | 76.3%        | 1,222    | 69.2%    |         |
|                               | Unknown  | 15,089       | 0.3%         | 46       | 2.6%     |         |
| SEN: Year 5                   | SEN      | 134,140      | 21.5%        | 342      | 22.0%    | 0.638   |
|                               | Non-SEN  | 489,279      | 78.5%        | 1,212    | 78.0%    |         |
|                               | Unknown  | 0            | —            | 211      | —        |         |
| Region: Year 5                | North    | 179,714      | 28.8%        | 972      | 55.1%    | < 0.001 |
|                               | Midlands | 194,777      | 31.2%        | 513      | 29.1%    |         |
|                               | South    | 248,928      | 39.9%        | 280      | 15.9%    |         |

- Sex and SEN were not significantly different from the national population ( $p = 0.183$  and  $p = 0.638$  respectively). The sample is representative on these dimensions.
- Pupil Premium eligibility was over-represented in the sample (31.4% vs 24.7% nationally,  $p < 0.001$ ).
- English as an Additional Language was over-represented in the sample (28.2% vs 23.4% nationally,  $p < 0.001$ ).
- The sample was geographically skewed towards the North of England (55.1% vs 28.8% nationally). Region is not strongly associated with attainment after other demographic factors are taken into account, and no weighting on region is applied (Manual §6.2).

Sample weights were applied during standardisation so that the effective sample matched the national population proportion for Pupil Premium eligibility (Manual §2.7.2). No other weights were applied.

## Reliability and SEM

The internal-consistency reliability of the assessment, measured by Cronbach's alpha, is **0.904** – above the target of 0.80 (Manual §3.1). This indicates that the items within the assessment work together coherently as a measure of the same underlying construct, with high consistency across the standardisation sample.

The Standard Error of Measurement is **4.6 standardised score points**. For a student with an observed standardised score of 100, the 95% confidence interval around their true score is approximately 91.0 to 109.0, and the 68% confidence interval is 95.4 to 104.6 (Manual §3.3). If the student were to sit comparable assessments many times, around 95 in every 100 of those scores would fall within the 95% range.

## Differential Item Functioning

Items in the Year 6 Baseline Mathematics assessment were examined for differential item functioning (DIF) by sex and by Pupil Premium eligibility, using a 2-PL multiple-group IRT model with Benjamini–Hochberg adjustment for multiple testing (Manual §3.4). The number of items showing statistically significant DIF is shown in Table 2.

**Table 2: Items showing significant DIF**

| Dimension     | Items with significant DIF | Total items | Proportion |
|---------------|----------------------------|-------------|------------|
| Sex           | 4                          | 28          | 14.3%      |
| Pupil Premium | 2                          | 28          | 7.1%       |

The magnitude of DIF for each flagged item was classified using the scale of Jodoin and Gierl (2001)<sup>12</sup> as either negligible or moderate. The items showing DIF were reviewed and no obvious patterns of bias were found; the flagged items spanned different curriculum domains, skills and item types.

## Item performance summary

All 28 items in the Year 6 Baseline Mathematics assessment passed the item-selection discard rules at the item-trial stage (Manual §2.6.4): each item had a facility within 10–90%, an omit-of-seen rate below 15%, an IRT slope above 0.4, and an IRT difficulty parameter within the range -3 to +3. The spread of facility values across the assessment supports measurement of attainment across the full ability range targeted by the Year 6 Baseline Mathematics Assessment (Manual §2.6.5).

<sup>12</sup> Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, 14(4), 329–349. [https://doi.org/10.1207/S15324818AME1404\\_2](https://doi.org/10.1207/S15324818AME1404_2)

The following summary statistics describe item performance in the standardisation sample (n = 1,765).

**Table 3: Item-level summary statistics**

| Item-level statistic             | Mean  | Range         |
|----------------------------------|-------|---------------|
| Facility (% correct)             | 46.0% | 19.1% – 91.5% |
| Corrected total-item correlation | 0.48  | 0.19 – 0.63   |
| Omission rate                    | 9.6%  | 0.7% – 21.7%  |
| Not-reached rate                 | 3.4%  | 0.0% – 8.9%   |

Items showing any concerning indicators in the standardisation sample (for example, a corrected total-item correlation below 0.3 or an omit-of-seen rate above 15%) are flagged for review at the next re-standardisation cycle (Manual §6.4).

## Year 6 Baseline Science

This supplement reports the technical results for the Year 6 Baseline Science assessment. It should be read alongside the CENTURY Academic Assessments Technical Manual, which sets out the methodology used to develop, trial, standardise and report all assessments in the suite. Section references in this document refer to that Manual.

### Assessment overview

The Year 6 Baseline Science assessment measures students' attainment at the start of Year 6 against the Year 5 Science Programme of Study of the English National Curriculum.

The assessment comprises **28 items**, designed to be completed within a single 50-minute sitting and covers the following topics: Animals including humans, Earth and space, Forces, Living things and their habitats, and Properties and changes of materials.

The Year 6 Baseline Science assessment is the autumn-term parallel form of the Year 5 End-of-Year Science content set. The two forms are equated parallel forms (Manual §1.2): items in each were drawn from a common trialled pool, calibrated to the same difficulty and balanced across the same curriculum domains. The Year 5 End-of-Year Science assessment is standardised separately during the summer term.

### Item trial

In the summer term of 2025, **96 items** were trialled with **722 Year 5 students** (pre-standardisation testing, which is detailed below) across six spiralled forms (Manual §2.6.1). The trialled item pool fed selection of items into both the Year 5 End-of-Year Science and Year 6 Baseline Science assessments. Classical item statistics and Item Response Theory (IRT) parameters were calculated for all trialled items, with all six trial forms analysed together (Manual §2.6.3). Of these, **28 items** were selected for the Year 6 Baseline Science assessment, balanced for curriculum domain coverage, mean IRT difficulty and spread of difficulty (Manual §2.6.5). The selected items achieved a marginal IRT reliability of **0.862** at final form construction.

### Standardisation

The assessment was standardised on **1,240 students** across 37 schools during September and October 2025 (the start of the academic year in England). Raw scores ranged from 0 to 28, with a mean of 14.5 (standard deviation 5.6) and a median of 15. The interquartile range (25th to 75th percentile) was 11 to 19.

The achieved sample composition was compared with national population figures published by the Department for Education using one-sample chi-square tests. Results are shown in Table 1.<sup>1</sup>

**Table 1: Sample composition vs national population**

| Dimension                     | Category | Population N | Population % | Sample N | Sample % | p-value |
|-------------------------------|----------|--------------|--------------|----------|----------|---------|
| Sex: Year 5                   | Male     | 316,321      | 50.7%        | 643      | 52.2%    | 0.294   |
|                               | Female   | 307,098      | 49.3%        | 588      | 47.8%    |         |
|                               | Unknown  | 0            | —            | 9        | —        |         |
| Pupil Premium: Primary sector | PP       | 1,126,353    | 24.7%        | 371      | 30.1%    | < 0.001 |
|                               | Non-PP   | 3,428,621    | 75.3%        | 863      | 69.9%    |         |
|                               | Unknown  | 0            | —            | 6        | —        |         |
| EAL: Primary sector           | EAL      | 1,065,551    | 23.4%        | 367      | 29.6%    | <0.001  |
|                               | Non-EAL  | 3,474,334    | 76.3%        | 866      | 69.8%    |         |
|                               | Unknown  | 15,089       | 0.3%         | 7        | 0.6%     |         |
| SEN: Year 5                   | SEN      | 134,140      | 21.5%        | 248      | 22.4%    | 0.493   |
|                               | Non-SEN  | 489,279      | 78.5%        | 861      | 77.6%    |         |
|                               | Unknown  | 0            | —            | 131      | —        |         |
| Region: Year 5                | North    | 179,714      | 28.8%        | 626      | 50.5%    | < 0.001 |
|                               | Midlands | 194,777      | 31.2%        | 361      | 29.1%    |         |
|                               | South    | 248,928      | 39.9%        | 253      | 20.4%    |         |

- Sex and SEN were not significantly different from the national population ( $p = 0.294$  and  $p = 0.493$  respectively). The sample is representative on these dimensions.
- Pupil Premium eligibility was over-represented in the sample (30.1% vs 24.7% nationally,  $p < 0.001$ ).
- English as an Additional Language was over-represented in the sample (29.6% vs 23.4% nationally,  $p < 0.001$ ).
- The sample was geographically skewed towards the North of England (50.5% vs 28.8% nationally). Region is not strongly associated with attainment after other demographic factors are taken into account, and no weighting on region is applied (Manual §6.2).

Sample weights were applied during standardisation so that the effective sample matched the national population proportion for Pupil Premium eligibility (Manual §2.7.2). No other weights were applied.

## Reliability and SEM

The internal-consistency reliability of the assessment, measured by Cronbach's alpha, is **0.844** – above the target of 0.80 (Manual §3.1). This indicates that the items within the assessment work together coherently as a measure of the same underlying construct, with high consistency across the standardisation sample.

The Standard Error of Measurement is **5.9 standardised score points**. For a student with an observed standardised score of 100, the 95% confidence interval around their true score is approximately 88.4 to 111.6, and the 68% confidence interval is 94.1 to 105.9 (Manual §3.3). If the student were to sit comparable assessments many times, around 95 in every 100 of those scores would fall within the 95% range.

## Differential Item Functioning

Items in the Year 6 Baseline Science assessment were examined for differential item functioning (DIF) by sex and by Pupil Premium eligibility, using a 2-PL multiple-group IRT model with Benjamini–Hochberg adjustment for multiple testing (Manual §3.4). The number of items showing statistically significant DIF is shown in Table 2.

**Table 2: Items showing significant DIF**

| Dimension     | Items with significant DIF | Total items | Proportion |
|---------------|----------------------------|-------------|------------|
| Sex           | 7                          | 28          | 25.0%      |
| Pupil Premium | 0                          | 28          | 0.0%       |

The magnitude of DIF for each flagged item was classified using the scale of Jodoin and Gierl (2001)<sup>13</sup> as negligible. The items showing DIF were reviewed and no obvious patterns of bias were found; the flagged items spanned different curriculum domains, skills and item types.

## Item performance summary

All 28 items in the Year 6 Baseline Science assessment passed the item-selection discard rules at the item-trial stage (Manual §2.6.4): each item had a facility within 10–90%, an omit-of-seen rate below 15%, an IRT slope above 0.4, and an IRT difficulty parameter within the range -3 to +3. The spread of facility values across the assessment supports measurement of attainment across the full ability range targeted by the Year 6 Baseline Science Assessment (Manual §2.6.5).

The following summary statistics describe item performance in the standardisation sample (n = 1,240).

<sup>13</sup> Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, 14(4), 329–349. [https://doi.org/10.1207/S15324818AME1404\\_2](https://doi.org/10.1207/S15324818AME1404_2)

**Table 3: Item-level summary statistics**

| Item-level statistic             | Mean  | Range         |
|----------------------------------|-------|---------------|
| Facility (% correct)             | 51.8% | 21.5% – 90.5% |
| Corrected total-item correlation | 37.0% | 0.17 – 0.59   |
| Omission rate                    | 6.8%  | 1.6% – 21.0%  |
| Not-reached rate                 | 2.9%  | 0.0% – 5.5%   |

Items showing any concerning indicators in the standardisation sample (for example, a corrected total-item correlation below 0.3 or an omit-of-seen rate above 15%) are flagged for review at the next re-standardisation cycle (Manual §6.4).