# Scottish Survey of Achievement 2008 Technical Annex

**Contents**                                                    **page**

# 1. Overview

## 1.1 Survey objectives

The 2008 Scottish Survey of Achievement (SSA) was required to:

- assess and report nationally on pupils' attainment in mathematics and numeracy by 5 to 14 levels at the P3, P5, P7 and S2 stages – referred to as "*understanding mathematics*"

- investigate and report nationally on pupils' ability to use numeracy skills in everyday contexts at the P3, P5, P7 and S2 stages – referred to as "*using mathematics*"

- assess and report nationally on pupils' core skills (ICT, working with others and problem solving) applied in a mathematics context at the P3, P5, P7 and S2 stages

- gather information and report nationally on pupils' and teachers' experience of learning and teaching in mathematics, along with their views about this experience

- gather and report nationally on teachers' judgements of pupils' reading, writing and mathematics attainment levels

- compare attainment and progression in mathematics and numeracy across the four stages and between girls and boys

- compare attainment and progression in mathematics and numeracy in relation to deprivation

- assess changes in performance over time where possible

and, as secondary objectives, to:
- assess and report, for 'opted-in' local authorities, pupils' attainment in mathematics and numeracy by 5 to 14 levels at the P3, P5, P7 and S2 stages

- moderate a national sample of class-based writing (in a mathematics context, where possible) at the P3, P5, P7 and S2 stages

- provide information to inform mathematics education policy and support opportunities for the enhancement of mathematics teaching in Scottish classrooms.  For example, by modelling new approaches to the assessment of mathematics in line with the Curriculum for Excellence.

The following practical constraints were imposed:

- the duration of an assessment session was to last no more than 30-40 minutes at P3/P5 and 50-60 minutes at P7/S2.
- the maximum that any individual pupil would be asked to undertake was two written booklets from either *understanding mathematics* or *using mathematics,* plus a questionnaire and *either* one of the three core skills elements of the practical together with a mental mathematics assessment *or* a piece of class-based writing
- the schools that had been invited to participate in the pre-testing of assessment material for the survey would not be selected for survey involvement, unless unavoidable
- the total number of pupils selected for testing in an individual school was to be in proportion to the size of the school roll, and was to average 13 for primary schools and 30 for secondary schools

- a maximum of twelve pupils per school were to be selected for participation in the practical elements of the survey.

## 1. 2  The scope of survey enquiry

### 1.2.1  Pencil and paper assessments

Pencil and paper assessment was employed for the assessment of *understanding mathematics* and *using mathematics*. In the first case, test booklets comprised conventional 'atomistic' test items, spanning three consecutive levels and covering the 5-14 mathematics curriculum for the stage concerned. A total of 120 different test booklets were administered, 30 per stage. Individual survey pupils were randomly allocated two booklets containing items at levels appropriate to the stage. Further details are given in Section 3. In the case of *using mathematics*, which was essentially a pilot study to inform future task development under Curriculum for Excellence, a handful of multi-item tasks were administered at each stage. These offered pupils an everyday context in which to demonstrate their ability to apply their numeracy skills. Again, individual survey pupils were randomly allocated two test booklets containing tasks at appropriate levels for the stage.

### 1.2.2  Practical assessments

As in previous years, the practical part of the SSA was carried out with a sub-sample of pupils in a sub-sample of the schools in the main survey (cost and logistics being too challenging for all of the survey schools to be included).  A target of 300 pupils per stage were required for each of the three elements of the practical assessments (working with others, ICT, maths investigation) as well as completing a mental maths assessment. Sampling was further constrained by the practicalities of using field officers to carry out the assessments. Although clearly not an ideal methodology from a statistical point of view, the main motivation for the practical assessments was to provide exemplification and professional development in practical assessment methodologies for schools. The information to be gathered was indicative only.

### 1.2.3  Writing

All survey schools were invited to be involved in this survey component, by submitting a piece of class-based writing for each of a stratified, randomly selected 25% of their survey pupils, along with a level judgement for the writing. Schools were given two options for selecting pupils' writing examples:
- a short or extended piece of functional writing in the context of mathematics, from the pupil's folio or generated specifically for the survey
- an extended piece of writing of any genre or context from the pupils' folio.
 A subset of submitted and rated writing was evaluated independently by teachers from other schools.

All useable submitted and rated writing was independently evaluated by a team of teachers recruited from across Scotland as part of an inter-rater study.

### 1.2.4 Teachers' level judgements

Continuing previous survey practice, the survey schools were invited to submit class teachers' level judgements in reading, writing and mathematics for each of their sampled pupils.

### 1.2.5 Teacher and pupil questionnaires

In order to provide a context against which to reflect on and interpret the attainment findings, both teachers and pupils were invited to complete questionnaires about their mathematics teaching/learning experiences and subject attitudes.

## 1.3 Reporting strategy

In May 2008, following the field work for the 2008 survey, the Scottish Government confirmed that the SSA provided statistics of national importance and as such should become a National Statistic. As a first step towards this the survey was designated as an "official statistic". This meant that, for the first time, the SSA would be fully managed within the National Statistics Code of Practice.

The practical implications included the following decisions about publication of 2008 survey findings:

- *understanding mathematics* (i.e. mathematics and numeracy attainment), teachers' judgements for mathematics, and questionnaire results would be published in the main 'headline' report;
- teachers' judgements for reading and writing would not be included in the headline report, but would be published within the supporting evidence;
- the *using mathematics* findings would not be published alongside the official statistics, but would be published later in another form, given its experimental nature looking forward to Curriculum for Excellence;
- the results of the practical investigations would also be reported separately, given their primary role in providing exemplification and professional development through the involvement of practising teachers as itinerant field officers;
- the results of the writing assessment would also be published separately, given their specific importance in informing future assessment practice in this area under Curriculum for Excellence.

For these reasons, the survey headline report includes selected findings from the principal assessment of mathematics and numeracy, from teachers' judgements, and from the teacher/pupil questionnaires. Other aspects of pupil performance assessed in the survey are reported separately.

The following sections document the sampling strategy used for the selection of school and pupils for survey participation, describes the strategy employed for the paper-based assessment of mathematics and numeracy (*understanding mathematics*), and overviews the methodology applied for attainment estimation.

# 2 School and pupil sampling

## 2.1 Overview

The principal aim of the 2008 SSA was to produce national estimates of attainment in mathematics and numeracy for pupils across Scotland at different stages in their education, whatever their medium of instruction (where a Gaelic medium learner was randomly selected, equivalent Gaelic language assessment material was made available). The only pupils deliberately excluded from the survey were those in special schools. Pupils with additional support needs who were being taught in mainstream schools could be withdrawn from the sample at the school's discretion, before or during testing, should their teacher or parent consider the experience potentially or actually distressing for them.

To achieve the required level of accuracy at a national level, a minimum of 3,600 pupils per stage would be selected to undertake the *understanding mathematics* assessment. In addition a further 1,500 pupils per stage were required to participate in the *using mathematics* assessment. As *using mathematics* was of secondary importance, and an experimental approach to measuring attainment looking forward to Curriculum for Excellence, a smaller sample was agreed to balance the burden on schools and pupils.

As in 2007, local authority Directors of Education were invited to opt their LA into authority level reporting of pupil attainment results for *understanding mathematics*, pupil and teacher questionnaire findings, and the profiles of teachers' judgements for mathematics, reading and writing. A total of 19 authorities chose to take advantage of this possibility: Aberdeen City, Angus, Argyll & Bute, Dumfries & Galloway, East Ayrshire, East Dunbartonshire, East Lothian, East Renfrewshire, Edinburgh, Falkirk (S2 only), Inverclyde, Midlothian, Moray, North Ayrshire, Perth & Kinross, Scottish Borders, South Ayrshire, South Lanarkshire, West Dunbartonshire. As in previous surveys, 450 pupils were to be selected at each relevant stage in each opted in authority for the assessment of *understanding mathematics* (in practice this number was increased slightly to allow for a 10% or so pupil loss through absence).

Taking into account these opt-ins, the intended total pupil sample size for the survey was just under 50,000 pupils. A two-stage disproportionate stratified random sampling scheme was applied to produce the national pupil sample for this principal mathematics/numeracy assessment. When national attainment estimates were calculated, the attainment data were duly weighted during analysis to address the over-representation of the reporting authorities in the national sample (see Section 4 for details); weighting was also applied when authority attainment estimates were produced, to address any gender or deprivation imbalances within authority samples. [The sampling strategy for *using mathematics* was proportionate stratified sampling, since there was no requirement to disproportionately increase authority samples for separate attainment reporting.]

## 2.2 School sampling

Within the primary sample every school would be asked to supply pupils at all three primary stages (P3, P5, P7) for assessment. For this reason a single school sample was drawn for each sector, primary and secondary. Prior to sampling taking place, authorities were given the opportunity to withdraw schools from the sampling frame in cases where it was considered difficult for them to participate. This occurred generally for logistic reasons – for example, the school building was closing during the year for refurbishment or staff shortages made administration extremely burdensome. Special schools were also excluded from the sampling frame. Schools

with both primary and secondary departments were treated as two schools (primary and secondary) for sampling purposes.

In both sectors, every sampled school was to supply pupils for the assessment both of *understanding mathematics* and of *using mathematics,* different pupils participating in one or other type of assessment. All schools were also to be invited to participate in the writing assessment component, with a subsample of schools also randomly selected for involvement in the practical assessments.

School sampling took place separately within each opted in authority, within the independent sector, and within the merged group of authorities that had not opted in for separate attainment reporting ("combined authority group").

In order to avoid excluding a handful of pupils from their classmates during testing, where a small school was selected to take part in the survey all the pupils in the stage were to be automatically included in the school's sample. For sampling purposes, publicly funded schools were therefore classified as "large" or "small": any school with fewer than ten pupils in a stage was classified as a "small" school[1]. In order to ensure that all the pupils within an authority (or within the combined authority group) had the same probability of selection, the school sampling probability was different for large and small schools. The total pupil requirement in each authority was split between the "large" and "small" schools on a pro-rata basis in terms of the proportion of pupils in the pupil population attending each type of school.

The numbers of small schools required in each authority, in the combined authority group and in the independent sector, were therefore determined by the total number of "small school" pupils needed for testing: schools were selected by simple random sampling until this number was reached.

The number of "large" schools to be selected was calculated by dividing the total required number of pupils in such schools in each stage in each authority by 30 at secondary and 13 at primary, these being the intended average numbers per school (in some reporting authorities this intention might not have been realised should there have been too few schools available in the school population). Simple random sampling was then used to select "large" schools in appropriate numbers for survey participation.

Selected schools were not compelled to participate in the survey, participation being at the discretion of the Head Teacher, and some schools did decline the invitation. Due to the timescales involved in survey planning and implementation there was no substitution on this occasion for schools that declined to take part. 85 per cent of sampled primary schools and 82 per cent of sampled secondary schools participated.

Pupils were selected for testing only from within those schools that agreed to participate. In order to maintain anonymity in the assessment materials and response data files, participating schools were allocated arbitrary but unique survey identifiers; when an "all through" school was selected in the primary and in the secondary school samples then the school was issued with different identifiers in the two sector samples.

## 2.3 Pupil sampling

Within an authority (or authority group) and stage, every pupil was to have an equal probability of inclusion in the sample. Multiplying the probability of a school being selected by the probability of a pupil within the school being selected would give this

---

[1] The total number of pupils in the stages of interest were based on the 2006 pupil census data, this being the latest available at this time in the process.

probability. In small schools all the pupils in the relevant stage(s) were automatically in the sample. This means that the within-school sampling fraction for pupils in small schools was 100%. At each stage, the within-school pupil sampling fraction for large schools varied from one authority to another, depending on the number of schools available in the sample, but was constant from one school to another within any particular authority, within the independent sector and within the authority group. This fraction was calculated for each stage in each stratum by taking the number of pupils required as a proportion of the total available pupils[2] in the "large" sampled schools in the relevant sector. This sampling fraction determined the required number of pupils to be randomly selected from within each selected large school. To minimize the burden on individual schools in the smaller authorities however, the aim was for the sampling fraction in their large schools to be no more than 50%. This issue was dealt with on an authority-by-authority basis, but in some cases the pupil sample size for the authority was reduced.

Where feasible, pupils in large schools were selected for involvement in the survey using stratified simple random sampling, the stratification variables being gender and deprivation category (full stratification was not feasible in single-sex schools or in schools with a preponderance of one deprivation category over the other). Individual pupils were classified into two deprivation categories based on the Scottish Index of Multiple Deprivation decile for their home postcode: "most deprived" (pupils living in any of the 20 per cent of datazones ranked as having "most" deprivation), and "less deprived". Where it was not possible to assign a SIMD decile to a pupil, because the pupil's home postcode was unavailable or non-valid, that pupil was allocated a category based on the postcode of their school. This occurred in about one per cent of cases. All pupils in independent schools were assumed to be in the less deprived category as no home address details are collected centrally. Pupils were selected using simple random sampling from within each stratum within each school, and issued with arbitrary unique survey identifiers.

The outcome of the sampling was an intended pupil sample for *understanding mathematics* of approximately 45,000 pupils in total, evenly distributed across P3, P5, P7 and S2. This represents approximately 20 per cent of the pupil population in these stages (or 22.5 per cent if *using mathematics* is counted). The pupils were drawn from just under 1,200 different schools throughout the country: 929 primary schools and 269 secondary schools. Detailed sample statistics are available in Chapter 8 of the 2008 supporting evidence.

Selected pupils were not compelled to participate in the survey. They could be withdrawn by their parents/carers or schools before the survey, or by their schools during testing. Pupils might also be absent during the testing period, or for one reason or another complete only one of their two randomly assigned test booklets (see Section 3). In the event, the overall pupil participation rate in this principal pencil and paper testing was over 85% in each sector.

## 2.4 Practical assessment sample

Around 100 schools per stage were selected to facilitate field officer visits. Four pupils per stage were allocated to each of the three elements (working with others, ICT, maths investigation) in every school. Pupils were selected from the main stage sample in each school by simple non-stratified random sampling. If there were fewer than twelve pupils per stage in a school then random allocation to assessments was constrained to ensure pupils were distributed across assessment types as equally as possible. To minimise the burden on individual schools, the P7 sample was taken from different schools to the P3 and P5 pupils.

---

[2] The pupil numbers were based on the 2007 pupil census which had become available by this time in the process.

# 3 The assessment materials and their administration

## 3.1 Defining mathematics and numeracy

The 2008 survey was required to report separate national attainment estimates for mathematics and numeracy. It therefore differed from the 2004 Assessment of Achievement survey, which reported on mathematics only, and the 2005 and 2006 SSA surveys which both reported on numeracy only.

The 5-14 Curriculum Guidelines define mathematics in terms of the following four attainment outcomes:

- Information handling
- Number, money and measurement
- Shape, position and movement
- Problem solving and enquiry

These outcomes are further defined in terms of strands representing important areas of skills development within the outcomes; for example, "add and subtract", "fractions, percentages and ratio", "symmetry" (see Table 3.1 for a complete list). Strands are then further defined in terms of attainment targets, which are specific learning outcomes at different levels of attainment, such as "use 12 hour times for simple timetables", "define and classify quadrilaterals", "describe the main features of a graph". Progression within outcomes and strands is defined in terms of six levels of attainment, A to F, with A being the lowest and F the highest.

**Table 3.1: The mathematics and numeracy outcomes and strands represented in the survey**

| |
| --- |
| *Information handling* |
| Interpret information * |
| *Number, Money & Measurement* |
| Range and type of numbers * |
| Money * |
| Add and subtract * |
| Multiply and divide * |
| Round numbers * |
| Fractions, percentages and ratio * |
| Patterns and sequences |
| Measure and estimate * |
| Time * |
| Perimeter, formulae and scales |
| Functions and equations |
| Algebra |
| *Shape, Position & Movement* |
| Range of shapes |
| Position and movement |
| Symmetry |
| Angle |
| *Problem solving and enquiry* |
| Problem solving |
| * "Numeracy" strands within mathematics |

There has never been an accepted definition of "numeracy" in the context of the 5-14 guidelines[3].The concept of numeracy used in the 2008 survey is the same as that used in the surveys of 2005 and 2006, and is based on the understanding of numeracy as a core skill within national qualifications. The focus is on computation skills and the interpretation of information presented in a variety of ways, in particular, but not exclusively, as tables, graphs or charts. Table 3.1 identifies those strands in the 5-14 mathematics guidelines which qualified as numeracy for this survey.

## 3.2 Test items

As in the earlier surveys, test booklets comprised atomistic test items drawn from the National Assessment Bank. Items at a given level were common across stages (for example, a Level B task used at P3 would also be used at P5) and a sufficient number of items from the 2006 SSA were repeated to ensure some comparison over time. Each item, classified by 5-14 curriculum strand and by level (one of A to F), typically invited a single short response to an instruction (subtract, multiply, etc), or to a question based on given information and in this way set "in context" (for examples see Figures 3.1 to 3.5). All items were dichotomously scored.

**Figure 3.1: Information handling (Level A)**
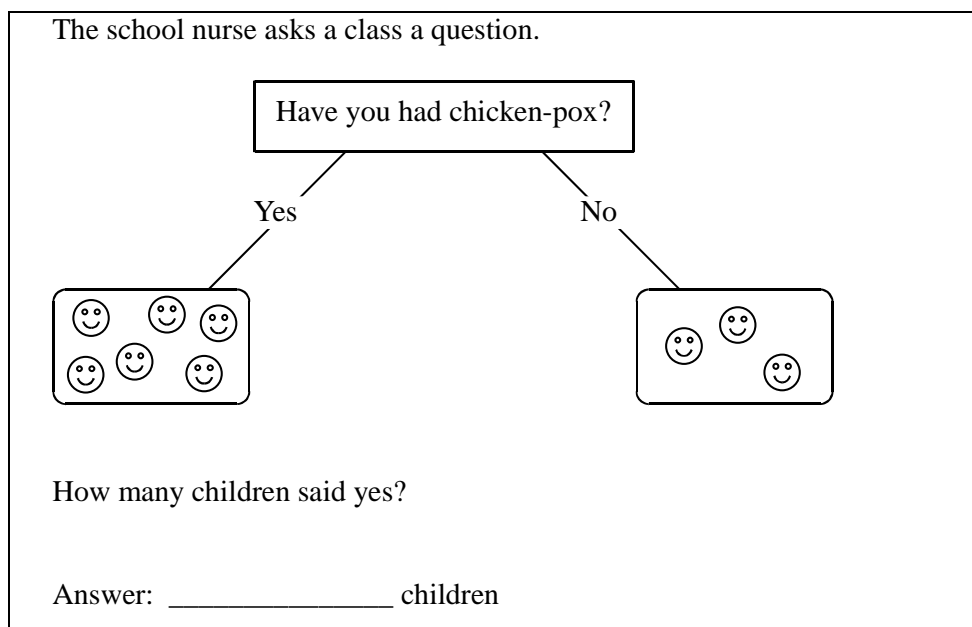


**Figure 3.2: Number, money & measurement (Level B)**
*(Multiplication item in context on the left, with an abstract subtraction item on the right)*

| There are 5 crayons in a packet.<br><br>How many crayons are there in 16 packets?<br><br><br><br>Answer: _____ crayons | Subtract<br><br>42 – 29<br><br><br><br>Answer: _____ |
|---|---|

---

**Figure 3.3: Problem solving (Level C)**

The 30 children in a P4 class each made a thumbprint.

There were three different kinds – whorl, loop and arch.

The same number of boys and girls had arch thumbprints.

Complete the table.

|  | **Whorl** | **Loop** | **Arch** |
|---|---|---|---|
| **Boys** | 8 | 5 |  |
| **Girls** | 4 | 5 |  |

**Figure 3.4: Problem solving (Level D)**

Bundles of concert programmes were given to five classes – A, B, C, D and E.

- There were enough programmes to give one to each child.
- Class A had the least number of children in the class.
- Each of the other classes had 1 more child on the register than the previous class.
- 150 programmes were handed out altogether.

How many children were in each class?
Write the answers in the register.

| Register | |
|---|---|
| Class A | _____ children |
| Class B | _____ children |
| Class C | _____ children |
| Class D | _____ children |
| Class E | _____ children |

**Figure 3.5: Number, money & measurement (Levels E and F)**
*(Level E addition in context on the left, Level F abstract division on the right)*

| | |
|---|---|
| Peter completes 4 stages of an obstacle course in the following times: <br><br>     10·65 seconds, 25·43 seconds, 18·57 seconds and 40·40 seconds. <br><br> What was Peter's total time to complete the course? <br><br><br><br> Answer: _____ seconds | Calculate <br><br>      $0·06 \div 0·3$ <br><br><br><br><br> Answer: _____ |

## 3.3 Test booklets

The principal objective of the survey was to assess and report on pupils' mathematics and numeracy skills at national and local authority level. Assessment was to be reported against two 5-14 levels at P3 (A and B) and three at the other stages (B, C and D at P5; C, D and E at P7; D, E and F at S2). The consequence of the requirement to report at both national *and* authority level on both mathematics *and* numeracy at two or three levels per stage, coupled with the constraint of a maximum of two test booklets per pupil, was to seriously increase the complexity of the survey design.

In previous AAP/SSA surveys of mathematics or numeracy, each pupil had taken two test booklets, each containing items at the relevant two or three levels. Items at the same level across the pair of booklets that any pupil was allocated comprised a test at that level. Thus, where booklets contained items at three levels, the pupil actually attempted three single-level tests. Figure 3.6 illustrates this. Pupil attainment was measured on each single-level test, and population attainment estimates then computed level by level.

**Figure 3.6: Relationship between test booklets and single-level tests**
**for numeracy at P5 in the 2006 survey**
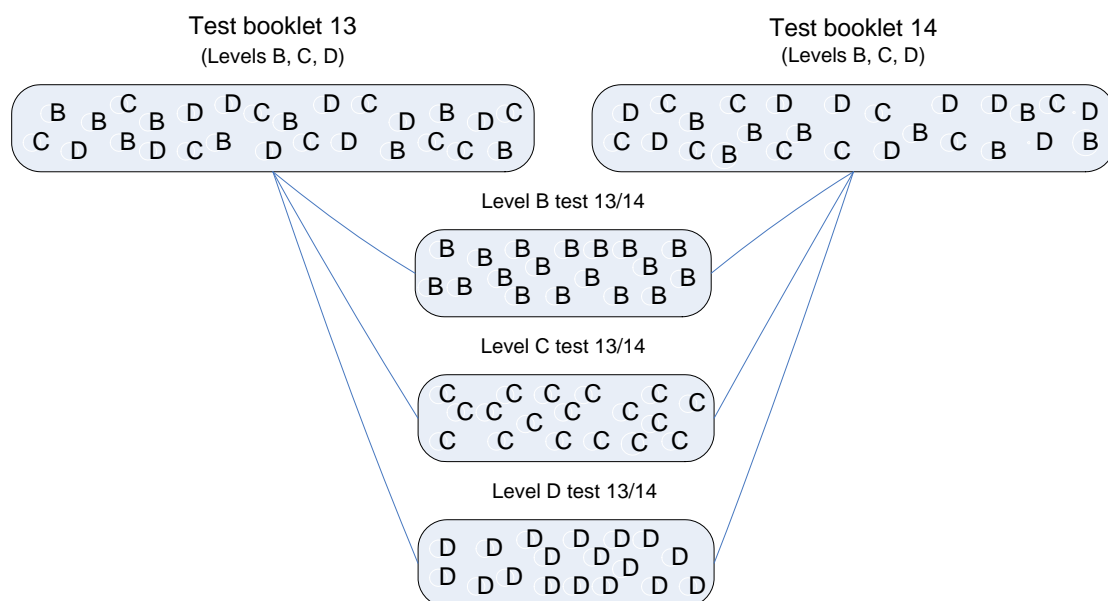
In the 2008 survey, it was not possible to include in two test booklets sufficient test items to permit attainment measurement for both mathematics and numeracy at every relevant level. It was decided, therefore, that booklets would need to be longer than before, and, in addition, that they would need to be designed in such a way that over the whole survey sample it would be possible to provide population attainment estimates for mathematics and for numeracy at all the levels required, even though this would not be possible for individual pupils.

In the event, three types of booklet were created for use at each stage. Each booklet type would enable pupil attainment measurement for mathematics and for numeracy, but only for two of the three relevant levels, and the levels concerned would be different for one aspect compared with the other. The whole set of items at one of the levels represented numeracy, as this featured in test booklets in the 2005/06 surveys. For a second level, the items taken together represented mathematics, as this featured in test booklets in the 2004 survey (i.e. subsuming numeracy items but not in numbers large enough to permit the separate measurement of numeracy skills). At the third level, the collection of items represented mathematics as this featured in 2004, but also numeracy, as this featured in 2005/06, the latter enabled through the addition of appropriate additional numbers of numeracy items at that level than would normally be included in a mathematics item set.

At P5, for example, the three booklet types were as follows in terms of their individual pupil attainment capability:

Type 1: level B (maths/numeracy) + level C (maths) + level D (numeracy)
Type 2: level C (maths/numeracy) + level D (maths) + level B (numeracy)
Type 3: level D (maths/numeracy) + level B (maths) + level C (numeracy)

Thus, a P5 pupil completing two Type 1 booklets would be assessed for both maths and numeracy separately at level B, for maths only at level C and for numeracy only at level D (see Figure 3.7). This same general pattern applied to the other stages, with respective changes in level combinations.

**Figure 3.7: Relationship between test booklets and single-level tests for mathematics and numeracy at P5 in the 2008 survey**
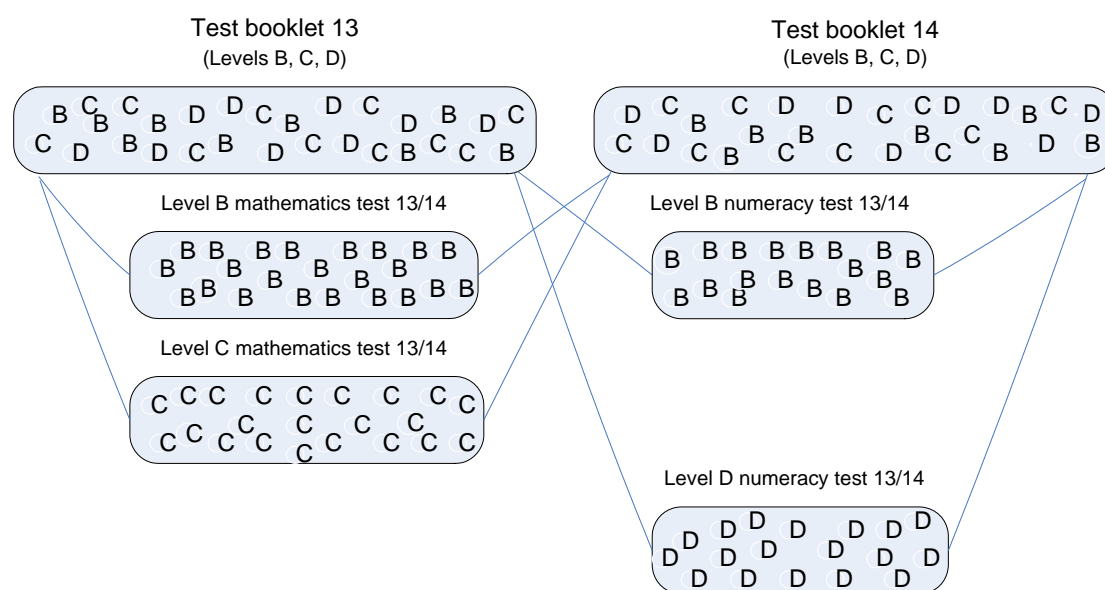


For each stage 10 test booklets were created of each type, these being similarly representative of the mathematics or numeracy curriculum at the appropriate levels, and therefore in principle interchangeable (no constraints were imposed to ensure equal "difficulty" or other empirical characteristics). The booklets were created

through a process of stratified random sampling of test items from within the National Assessment Bank, the stratification being determined by the associated test specification. Once selected, the items to comprise a particular booklet were randomly ordered for presentation, becoming booklet version "A". The order of presentation was then simply reversed to create booklet version "B". This common strategy was to minimise the unwanted effects of test fatigue on individual item statistics.

In total, therefore, 120 different test booklets were created for use in the survey, at 30 per stage, each booklet produced with two different item presentation orders. Items at the same level in one booklet type at one stage were repeated in a booklet of the same general type at other relevant stages. Thus, for example, all level B items in a Type 1 booklet at P3 (where they were mixed with level A items) were repeated in a Type 1 booklet at P5 (where they were mixed with items at level C and D).

Within each booklet the number of items at a level representing each of mathematics and numeracy varied between nine and eleven, depending on the level, while the combination mathematics/numeracy was represented by 13 or 14 items. Over two booklets the number of items constituting a single-level test therefore varied from 18 (for example for numeracy at level B) to 22 (mathematics at level F). At P3 a small number of additional level C items were added to each booklet, for later item by item reporting to inform development of Curriculum for Excellence. These were drawn from the outcomes Shape, Position & Movement and Problem solving, and from those strands in Number, Money & Measurement that are outside the survey definition of numeracy. As a result, booklets were 25 items long at P3 rising to between 33 and 35 items at the older stages.

## 3.4 Booklet administration

Test booklets were allocated to survey pupils using multiple matrix sampling, following the spiral design shown in Table 3.2. This strategy ensures that as many test items as possible are used in a survey, maximising curriculum coverage and hence assessment validity, without any one pupil being required to attempt unacceptably long tests, or to be assessed over unacceptably long periods of time. It also ensures that individual test items are as validly and reliably assessed for performance as possible. Booklets were randomly allocated to pupils in such a way that as few pupils as possible would be faced with the same booklet in any particular school, thus minimising any possibility of school effects. And all booklets, and their constituent items, would be attempted by similarly sized and similarly representative national (and authority) samples of pupils.

Assessment sessions were organised by the schools themselves, during May and June, with teachers supervising the testing. Schools were advised to arrange two separate assessment sessions for any group of pupils, to allow for a break between booklets. The sessions could be carried out on the same day or a day or two apart. Teachers were advised to allow 40 minutes (at P3 and P5) and 60 minutes (at P7 and S2) for each assessment session, but, at the schools' discretion, pupils could be given as much time as necessary to complete the tests. Pupils were not allowed the use of a calculator. Completed test booklets were returned from schools for pre-analysis processing.

**Table 3.2: The booklet allocation strategy at P5**

| Pupil | Booklets |
|-------|----------|
| 1 | 31A+32B |
| 2 | 31B+32A |
| 3 | 32A+33B |
| 4 | 32B+33A |
| 5 | 33A+34B |
| 6 | 33B+34A |
| 7 | 34A+35B |
| 8 | 34B+35A |
| 9 | 35A+36B |
| 10 | 35B+36A |
| 11 | 36A+37B |
| 12 | 36B+37A |
| 13 | 37A+38B |
| 14 | 37B+38A |
| 15 | 38A+39B |
| 16 | 38B+39A |
| 17 | 39A+40B |
| 18 | 39B+40A |
| 19 | 40A+31B |
| 20 | 40B+31A |

# 4. Population attainment estimation

## 4.1 Response processing

Pupils' responses were transcribed onto specially designed recording sheets, which listed a number of numeric answers, letter codes or key terms for each item (see Figure 4.1). For a multiple choice item, for example, the response option codes would simply be the letter codes associated with the displayed response choices, "A, B, C, D, E" perhaps. In other cases the response options might be numbers, as in cases where the item required the pupil to calculate a value (as in the item examples given in Figures 3.1 to 3.5). In still other, relatively rare, cases the response options might be phrases, drawings or key terms. Incorrect response possibilities were coded along with correct responses, where the incorrect responses might carry diagnostic value. Transcribers simply shaded the box alongside the response option(s) that matched the pupil's response.

**Figure 4.1: Extract (reduced size) from a booklet recording sheet**

| SSA 2008 - Understanding Mathematics | | M086B P7 | School [][][] | Pupil [][][] | Coder [][] | M086B P7 | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 [] Blank | 1 [] correct pattern | 2 [] | 3 [] | 4 [] | 5 [] | 6 [] | 9 [] Other |
| 2 | 0 [] Blank | 1 [] 140 | 2 [] 860 | 3 [] | 4 [] | 5 [] | 6 [] | 9 [] Other |
| 3 | 0 [] Blank | 1 [] 80.94 | 2 [] 8094 | 3 [] | 4 [] | 5 [] | 6 [] | 9 [] Other |
| 4 | 0 [] Blank | 1 [] number of programmes x25p | 2 [] x25(p) | 3 [] 25p for every programme sold | 4 [] | 5 [] | 6 [] | 9 [] Other |
| 5 | 0 [] Blank | 1 [] 12.15(pm) - 2.45(pm) | 2 [] | 3 [] | 4 [] | 5 [] | 6 [] | 9 [] Other |
| 6 | 0 [] Blank | 1 [] 5 | 2 [] 10 | 3 [] | 4 [] | 5 [] | 6 [] | 9 [] Other |
| 7 | 0 [] Blank | 1 [] 13.2 | 2 [] 23.5 | 3 [] | 4 [] | 5 [] | 6 [] | 9 [] Other |
| 8 | 0 [] Blank | 1 [] 1250 | 2 [] 1000 | 3 [] 1200 | 4 [] 1300 | 5 [] | 6 [] | 9 [] Other |
| 9 | 0 [] Blank | 1 [] 210 | 2 [] | 3 [] | 4 [] | 5 [] | 6 [] | 9 [] Other |
| 10 | 0 [] Blank | 1 [] 36 | 2 [] | 3 [] | 4 [] | 5 [] | 6 [] | 9 [] Other |

This transcription (or "coding") process was carried out by clerical staff, but supervised by subject specialists to ensure consistency and to mediate on any problematic judgements. One in ten of each booklet was independently re-coded, the two coding sheets compared and any differences highlighted. The supervising expert reviewed the differences and adjudicated as to whether the difference was due to a coder error or a re-coder error. This allowed a check to be carried out on the quality of the coding by individuals, and also alerted the supervisor to any potential problems with the coding of particular questions. The data from the coding sheets were keyed by a commercial data bureau for later automatic marking.

All the test items were dichotomously scored. In the majority of cases there would be a single correct answer meriting a mark. In other cases the mark could be gained from any one of a number of alternative types of response, or from some combination of responses. It remained to identify (in the recorded response data for each test booklet) the response, alternative responses, or combination of responses, that qualified the pupil for the mark for a particular test item, and thus to allocate the mark.

Once all the item level responses had been processed in this way, test scores were produced for pupils who had completed their two allocated booklets, for each single-level test that their booklets included (the booklet allocation strategy described in Section 3 resulted in 40 different single-level tests for each level assessed at a stage, for each of mathematics and numeracy). Cut-off scores were applied to the pupil test scores associated with each level test, in either mathematics or numeracy, and pupils

classified accordingly as 'good start' (50% or more of the items correct), 'well established' (65% or more of the items correct) or 'very good' (80% or more of the items correct) at the level. These criteria were first agreed on the basis of subject specialist judgement ahead of the 2001 AAP English Language survey, and have been systematically used in AAP/SSA surveys since then.

## 4.2  Data weighting

Estimated proportions of pupils in each attainment band in the population as a whole were calculated using data weighting, to address any gender and/or deprivation imbalances in the authority samples (for authority estimates) and the over-representation of opted-in authorities in the national sample at each stage (for national estimates).

The weighting attached to each pupil comprised two components. The first part of the weighting adjusts for imbalances in the pupil sample within the school, and is equal to the total number of pupils in the school who are in the same stage and have the same gender and deprivation classification as the pupil, divided by the number of those pupils who were included in the assessment.  The second part of the weighting adjusts for imbalances at the authority level and is equal to the total number of pupils in the authority with the same gender, deprivation classification and stage as the pupil, divided by the total number of such pupils who attended a school that participated in the assessments.

Applying these weights in turn produces estimates of the average population attainment of pupils in the stage, and with the given gender and deprivation category, or, multiplying by 100, the estimated percentage of pupils in that subgroup population who could be said to have attained the level. Estimates over larger subgroups, for example "all boys" as opposed to "most deprived" boys and "less deprived boys", and for the population as a whole, were produced by calculating weighted averages (or percentages).

## 4.3  Sampling error estimation

The pupil sample was selected using a complex multi-stage sampling technique, which means that the standard formulas used to calculate the standard error from a simple random sample would not be appropriate. Standard errors were therefore calculated empirically, using the jackknife procedure. The jackknife procedure is often referred to as the "leave one out" method. The principle is that, given a dataset with n observations (or sampling units), n re-sampled datasets can be created by excluding each observation in turn from the original dataset. Variability among the n new datasets allows us to calculate an unbiased estimate of the standard error of measurement for the attainment estimates based on the original dataset.