Document Title: Long scales for AY 2018-19: Briefing paper

DOCUMENT CONTROL

PRODUCT TITLE	Long scales for AY 2018-19
PRODUCT CODE	AR01
DOCUMENT No.	AR01-11

VERSION HISTORY

Date	Version	Comments	Author	Approved by	Date Approved
13/05/18	0.1	Initial draft	Redacted - exemption applies		18/05/18

Contents

1 Introduction

At the inception of work on developing the SNSA, Scottish Government and ACER agreed that in the first year of implementation (AY 2017-18) an interim reporting scheme would be used, based on an international scale for each subject area developed by ACER that reports learning in the subject area on a continuum, from early years to mid-secondary school. It was also agreed that, from the second year of implementation (AY 2018-19), SNSA would be reported on a new scale for each subject area, constructed based on data from Scottish children and young people. Each of these scales can be used to report capacities from early years to mid-secondary school, thus enabling monitoring of progress over time. This paper outlines how the Scottish long scales have been developed, and the proposed school-level reporting for AY 2018-19.

In designing the reports for AY 2018-19, care has been taken to preserve the formats of AY 2017-18 reporting as far as possible, in order to allow teachers, school leaders and local authorities to build on their understanding of the features of SNSA reporting that they have already assimilated. At the same time the new reporting will allow more nuanced interpretation of children's and young people's literacy and numeracy capacities and, over time, the ability to monitor trends at individual, class and school levels.

The basic suite of school-level reports under development for the beginning of AY 2018-19 is similar to that for AY 2018-19:

- Individual learner report
- Group diagnostic report
- Group aggregate report

Later in AY 2018-19, once equating with the AY 2017-18 scale (the interim scheme) has been completed, results for the first year of implementation will be transposed onto the Scottish scale, and a year-on-year trend report, comparing year group performance at class or school level, will be available.

Local authority reports for AY 2018-19 will also be reported on the long scales. These will be discussed at a later stage.

This paper first outlines the key stages in the construction of the long scales for Scotland, and then presents mock-ups of the school-based reports under development.

2 Construction of the Scottish long scales

Data used for constructing the long scales described in this paper are from the November 2017 and March 2018 norming studies, and from the long scale equating study conducted in February–March 2018.

The SNSA comprises 11 assessments, at four stages, in three subject areas, as shown in Table 1.

Table 1 Assessments administered in SNSA AY 2017-18

Primary 1	literacy	numeracy	
Primary 4	reading	numeracy	writing
Primary 7	reading	numeracy	writing
Secondary 3	reading	numeracy	writing

The scales are constructed using data from learner's completion of assessments in each subject area. Each subject area has assessments that measure capacity in that particular area; therefore each subject area has its own separate scale.

The aim of long scale development for SNSA is to make it possible to measure capacities of learners in different stages of a subject area on the same scale, thus allowing monitoring of growth as the learners move through the stages. Three scales have been constructed: one for reading, for the reporting of P1(literacy), P4 reading, P7 reading and S3 reading capacities; one for numeracy, for the reporting of P1, P4, P7 and S3 numeracy capacities; and one for writing, for reporting P4, P7 and S3 writing capacities.

There are two basic ingredients for a robust reporting scale. First, each assessment must comprise items that are psychometrically robust, and the set of items as a whole must measure a unified domain (e.g. numeracy). Secondly, to construct a long scale, there must be sound statistical linkage between the sets of items in each subject area's assessment. The extent to which SNSA meets these criteria is addressed below.

2.1 Robustness of items within each assessment

To evaluate whether a set of items is psychometrically sound, it is necessary that each item discriminates effectively between learners who have the knowledge or skill that underpins the item, and those who have not; and the set of items as a whole measures a unified latent trait (e.g. numeracy). One key summary statistic that is usually used to indicate that items in an assessment are working well as a set is a reliability statistic. Table 2 shows the reliabilities of the 11 assessments using Cronbach's Alphas. Reliabilities above 0.75 are typically considered as satisfactory and above 0.80 as excellent.

Table 2 Assessments' reliability results SNSA AY 2017-18

Form	Reliability
P1 numeracy	0.840
P1 literacy	0.849
P4 numeracy	0.868
P4 reading	0.880
P4 writing	0.882
P7 numeracy	0.889
P7 reading	0.860
P7 writing	0.820
S3 numeracy	0.880
S3 reading	0.887
S3 writing	0.780

The summary of the results presented in Table 2 shows that the reliabilities of all assessments forms were satisfactory to excellent (more than 0.75).

2.2 Linking between assessments

Once the validity of each assessment has been established, the next step is to equate the assessments across year groups, to place them on the same scale. With a wide gap of three school years between each of the assessments in a subject area (P1, P4, P7 and S3), the most efficient way of equating is to administer assessments to learners in intervening year groups that contain items from the year group assessments on either side. Accordingly, a 'long scale equating study' was implemented in February and early March 2018 to a sample of Scottish children and young people in P2, P3, P5, P6, S1 and S2. Each of the assessments was constructed using clusters of items from adjacent and adjacent-but-one year groups. For example, the P2 and P3 numeracy assessments both included P1 and P4 numeracy items, with more P1 items in the P2 assessment, and more P4 items in the P3 assessment. A similar pattern was followed throughout.

The methodology used to evaluate the quality of the linking (and therefore validating the equating to create a long common scale) is to check whether the items are behaving in a similar, parallel fashion with the different year groups. Ideally, for example, children in P3 will succeed relatively less often on P4 numeracy items than children in P4, and to a similar degree. (A difficult item for P4s will be even more difficult for P3s; an easy item for P4s will be relatively easy for P3s, but not as easy as for P4s.) Items that are shown to be dissimilar in their performance in different year groups are not considered good linking items and would not be used for the equating. On the whole the long scale equating study demonstrated reasonably good linking between assessments in adjacent year groups.

For the equating analysis, performance of the linked items was compared with their performance in the adjacent year (e.g. P1 item difficulties with those of the same items in P2 and P3, item difficulties in P2 and P3 with those of the same items in P4, etc.) Items that were not performing well as links were removed from the equating. With the resulting final sets of linked items, the differences in mean difficulties were used to determine the locations of items from these different assessments to be put on the common scale.

The equating from P4 to S3 is unproblematic. More items were removed from linking between P1 and P2 literacy, and between P1 and P2 numeracy, than from the other linked sets of items. These kinds of results have been observed in other contexts where the aim has been to map the skills

development of very young children onto a scale with those of children in higher years of schooling. [1] Appendix 1 provides more detailed illustration of the issue.

Nevertheless, we recommend that all three subject areas be reported on long scales, from P1 to S3 in the case of reading and numeracy, and from P4 to S3 in the case of writing.

'Don't let the perfect be the enemy of the good.'

While there is some degree of uncertainty in the linkage from P1 to P4 for both literacy and numeracy, our recommendation is to build the long scale and its reporting in AY 2018-19 on the best estimate we have of that linkage, and to accompany the reporting with indications that the equating as a whole is provisional, and will be reviewed and adjusted if necessary over the coming two years.

The recommendation is made taking into account the fact that SNSA's mapping of individuals' and cohorts' growth between P1 and P4 will not be possible until AY 2020-21, when the children who were in P1 in AY 2017-18 reach P4. This delay will allow two more years of collection and review of data to interrogate the stability of the linking between stages. A research and analysis plan can be developed if SG wishes to go ahead with this methodology. Maintaining continuity in the framework and construction of the item pools – particularly at P1, but more generally for all of the assessments – will be an important ingredient for maintaining the stability of the measuring scale.

'If you want to measure change, don't change the measure.'

If the linking proves to be stable over this period, the long scales commencing at P1 for literacy/reading and for numeracy may validly be confirmed. If not, there would be several options, including

- limiting the P1 assessments to items that work well in fitting to the long scale (e.g. for literacy, only including comprehension items;
- reporting P1 results on a stand-alone scale (rather than the long scale), which would still allow year-on-year trend analysis (P1 2017-18 compared with P1 2018-19, P1 in 2019-20, etc.); or
- providing only diagnostic, item-by-item reporting on P1 assessments.

[1] There are a number of possible explanations for this, and they differ for literacy and numeracy.

One of the contributing factors is that growth in learning at this emerging stage is rapid and may be sudden, in spurts, rather than on a smooth course as the model assumes. In the case of literacy, in particular, the assessment comprises two distinct kinds of tasks: those that measure what Scott Paris has called 'constrained skills' – skills that are learnt and mastered completely by most children within the first one or two years of schooling; and 'unconstrained skills' – skills that continue to develop over time. An example of a constrained skill is knowledge of the letters of the alphabet. An example of an unconstrained skill is text comprehension. Because something like alphabet knowledge is legitimately assessed at P1, but is irrelevant for most children at P4 (and for many at P2), items addressing that knowledge do not show parallel difficulty for the two year groups. Our analysis shows that, of the items that do not 'fit' between P1 and P2, the majority are measuring these 'constrained skills'.

In the case of numeracy, items assessing content beyond early level curriculum were found to be relatively more difficult for children in P1 than in P2&3, which is not unexpected: both norming studies, on which the P1 data are based, were conducted too early in the school year for even the more capable P1 children to have encountered this advanced content. Other areas, where items were found to be relatively easy for children in P1, included areas where everyday experience (in contrast with curriculum experience) will help to develop understanding, e.g. recognition of the new £1 coin and reading analogue clock times.

2.3 Distribution of Scottish learner capacities across the calibrated long scales

125 The ensuing discussion assumes that in AY 2018-19 all three subject areas will be reported on long scales, from P1 to S3 in the case of reading and numeracy, and from P4 to S3 in the case of writing.

Figure 1 shows how each year group's capacity in numeracy, reading (literacy) and writing is distributed, once the results of the assessments have been calibrated onto a single long scale. The red diamond is the mean of each year group's capacity; the dark green horizontal bar is the median;
130 the dark blue shading is the middle 50%, and the mid-blue is the 5th percentile to 95 percentile.[2]

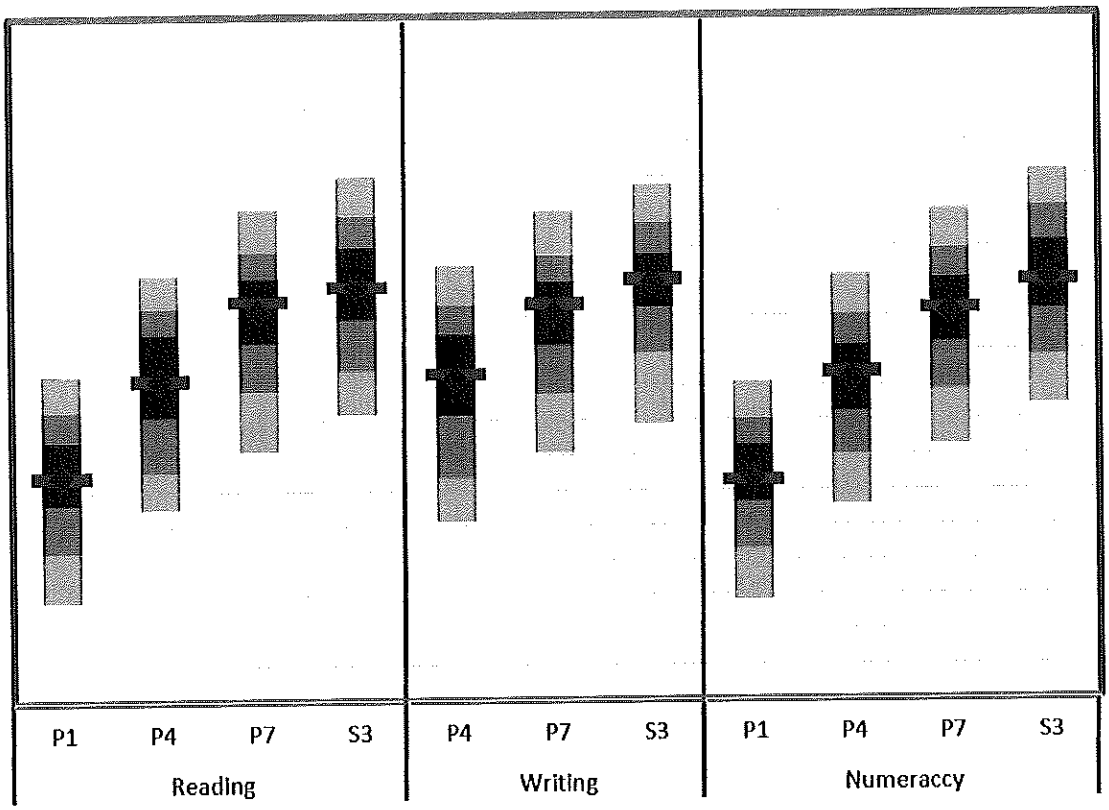


Figure 1 Sketch of progression from P1 to S3

Figure 1 shows that there is much overlap in capacity between the year groups, increasingly so as the year groups get higher.

2.4 Metric for the long scales

135 The metric proposed for the three long scales – reading, numeracy and writing – takes the mean capacity of P4 in AY 2017-18 as the centre of the calibration.

In order to translate the scale into a more interpretable form than scores, bands of capacity are drawn across the scale, above and below the mid-point, with a range of scores in each band. The

[2] Please note that this is a sketch only, with all three subject areas presented on the same figure for simplicity's sake. The results for numeracy, reading and writing do not sit on the same scale.

bands are numbered, arbitrarily, from 1 to n (depending on the number of bands), and described in words.

The logit scale (the basic unit of measurement used in item response theory) is converted to a more user-friendly metric, with a scale score of 500 set for the P4 mean in AY 2017-18. The variance across subject areas differs, so the standard deviations differ across subject areas. However, for each subject area a unique algorithm has been used to convert the logits to scale scores. In this way, the number of bands in the proposed model has been standardised, such that each year group can be reported on the same range of bands, as shown below in Figure 2.

With
RP=0.667

	Numeracy				Reading (literacy)				Writing		
BAND	P1	P4	P7	S3	P1	P4	P7	S3	P4	P7	S3
12											
11											
10											
9											
8											
7											
6											
5											
4											
3											
2											
1											

Figure 2 Reporting bands for each year group, for all subject areas

This design has the following features.

- The metrics for the three subject areas have the same anchor point: P4 mean in AY 2017-18, set at a scale score of 500.
- The number of bands is the same for numeracy and reading (but fewer for writing, as the scale only spans P4 to S3).
- Each year group will be reported on six bands.
- The range of bands on which results will be reported per year group will be the same across subject areas, with the same range of score points in each band.
- The lowest scale score will be greater than zero; the highest scale score will be below 1000. (There is a good buffer at both ends of the scale, but if capacity improves remarkably, 1000 may be exceeded.)

- The metric is refined enough to be likely to show growth over time in the vast majority of learners.
- Overall, each band encompasses sufficient items that have been administered to Scottish learners to give examples for the generalised description of the skills, knowledge and understanding that they address.

Except for the top and bottom of the scales, bands appear on the reports for more than one year group. This is because of the overlap of capacities, as seen in Figure 1. While the bands have the same numerical value across year groups, and indicate the same degree of capacity as a 'latent trait', the descriptions of the bands will vary slightly across year groups. This is to accommodate the fact that the manifestation of that trait will vary according the way it expresses itself at different stages (as reflected in the different items presented in each year group's assessment).

As agreed for AY 2017-18, the response probability (RP) value is 66.66% (RP66). This means that reporting a child or young person as demonstrating capacity of 'Band xx', means that the child or young person typically succeeds on items within that band 66% of the time – in other words, the child or young person is *twice as likely as not* to succeed on tasks of the Band XX difficulty level.

2.5 Described bands of capacity

As mentioned in the previous section, the long scale reports will include descriptions of what it means for a child or young person to achieve a band. This reporting allows interpretation of the bands and scale scores. The descriptions will initially be based on the pool of items included in the AY 2017-18 assessments, together with items that have been successfully trial tested with Scottish learners, through the in-test trialling conducted in 2017-18.

The SNSA is designed to reflect the Curriculum for Excellence (CfE), and pools of items selected for the assessments have aimed from the outset to address parts of CfE as they are articulated in the literacy and numeracy benchmarks and organisers. SNSA band descriptions will align with the content and language of CfE as far as possible, while also reflecting empirical evidence from the assessment items. As SNSA matures, the item pool will be refreshed and replenished to increase its approximation to the CfE. As a concomitant, the band descriptions will be reviewed and modified over time with a view to bringing them ever closer to the concepts and content of CfE.

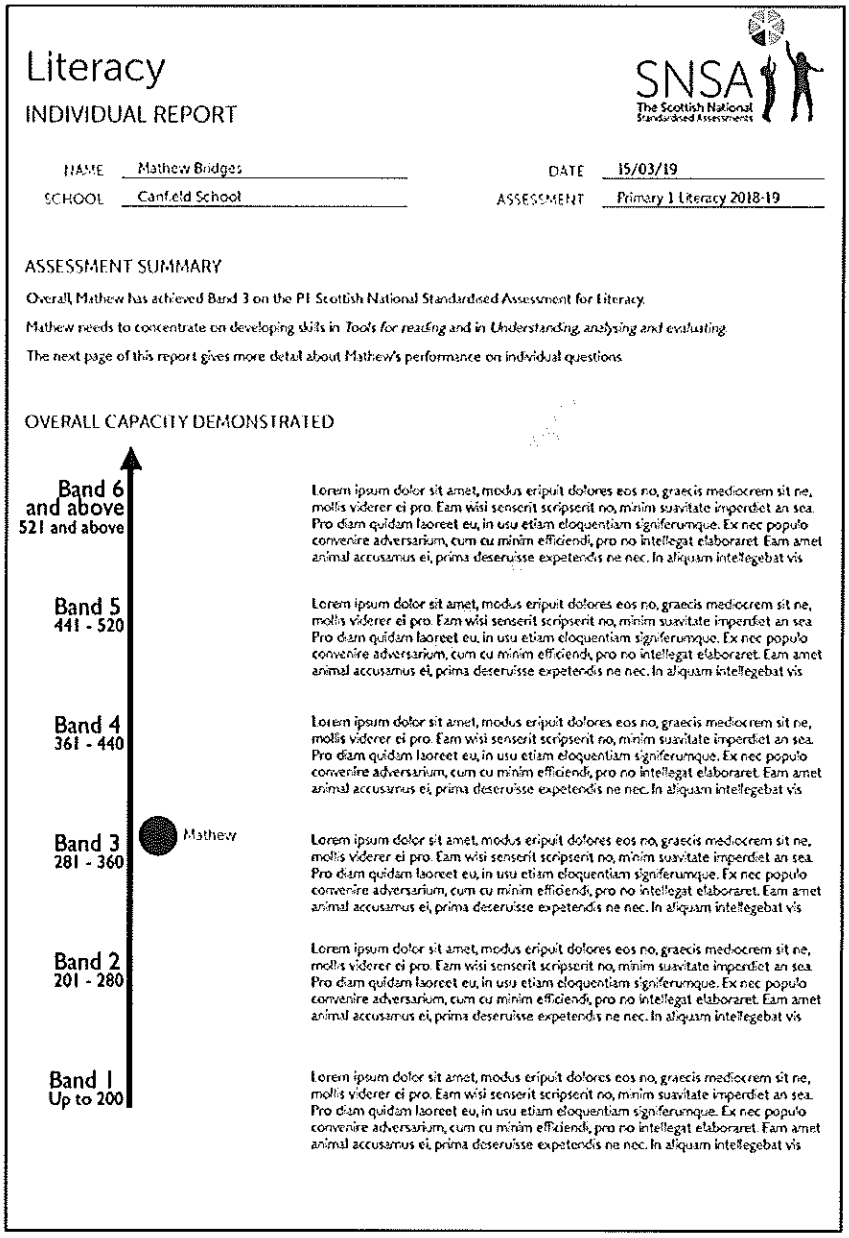
3 School level reports scheduled for release at beginning of AY 2017-18

This section provides mock-ups of the school-level reports. Please note that the mock-ups are in draft form: text description following each mock-up indicates where the mock-ups differ from what is planned for the live reports.

Please note also that the scale score ranges shown in the mock-ups are draft only.

3.1 Individual report

3.1.1 Section 1 mock-up and description



The layout should be familiar to users, since it is quite similar to the layout of the 2017-18 individual report. The Assessment summary refers to the band in which the learner's capacity lies, rather than

High, Medium or Low. The organiser descriptions (e.g. 'Mathew needs to concentrate on developing skills in Tools for Reading ...') will be similar to the ones for AY 2017-18. The graphic display will include the scale to the left, represented by a line with arrows. For P1 (as in the mock-up illustration) the arrow only appears at the top. Each band label has a scale score range printed below it. The child's capacity is represented by a circle. A generalised description of the skills, knowledge and understanding typically demonstrated by a learner at the given capacity is to the right of the graphic. The descriptions will be based on the items that have been presented in the SNSA in this region of the scale for this year group (though not necessarily on the items that this learner has seen).

Six band descriptions will be shown on each report. This P1 report shows Bands 1 to 6. P4 reports will show Bands 4 to 9; P7 reports will show Bands 6 to 11 and S3 reports will show Bands 7 to 12. (See Figure 2 above.)

If a learner's result is outside the 'relevant' range of bands for the assessment, their result will be reported as the highest or lowest band in that year group's set, with 'and above' or 'and below' on the band label (except for Band 1 where there will be no 'and below'). For example, since the highest band shown on the P1 reports is Band 6, the label says 'Band 6 and above'. The capacity of a P1 child performing above Band 6 would be shown in a circle at the top of the scale in section 1 and section 3 of the individual report. Labels on other reports will say 'Band 9 and above' etc. The child's actual scale score will be shown where relevant.

3.1.2 Section 2 mock-up and description

Literacy

INDIVIDUAL REPORT

SNSA The Scottish National Standardised Assessments

NAME Mathew Bridges DATE 15/03/19

SCHOOL Canfield School ASSESSMENT Primary 1 Literacy 2018-19

TOOLS FOR READING

Descriptor	Difficulty	Result
Identify a high frequency, single-letter word, in a sentence, with no audio support	Band 3	✘ Incorrect
Identify a synonym for a word in the context of a short narrative with audio support	Band 3	✔ Correct
Identify the text that shows the title on a book cover	Band 4	✘ Incorrect
Identify a common digraph at the end of a word	Band 3	✘ Incorrect
Identify the first sound in a word differentiating between similar sounding letters	Band 4	✔ Correct
Match a word to a picture when all the words start with /p/,without audio support	Band 2	Not Attempted
Match a picture to a word when all the images start with /r/without audio support	Band 2	✔ Correct
Match a word to a picture when all the words start with /h/ without audio support	Band 1	✘ Incorrect
Identify the first letter of a short phonetically, regular word	Band 2	✔ Correct
Identify the first letter of a short phonetically, regular word	Band 4	✘ Incorrect
Match a letter to its sound	Band 2	✘ Incorrect
Match a picture to a word when images start with different sounds	Band 2	✘ Incorrect
Distinguish a letter from a picture, a symbol and a numeral	Band 1	✘ Incorrect
Predict a narrative's content from viewing a book cover	Band 1	✔ Correct
Identify the last letter of a 3-letter, phonetically regular word	Band 6 and above	Not Attempted
Identify the upper case version of a lower case letter	Band 2	Not Attempted
Replace the first letter of a 3-letter, phonetically regular word to change its meaning	Band 2	Not Attempted
Match a letter to its sound	Band 2	✘ Incorrect
Match a picture to a word when the images start with different letters	Band 2	✔ Correct
Distinguish a word from a symbol and an image	Band 2	✘ Incorrect
Match a letter to its sound	Band 6 and above	✘ Incorrect
Match a letter to its sound	Band 1	✔ Correct

UNDERSTANDING, ANALYSING AND EVALUATING

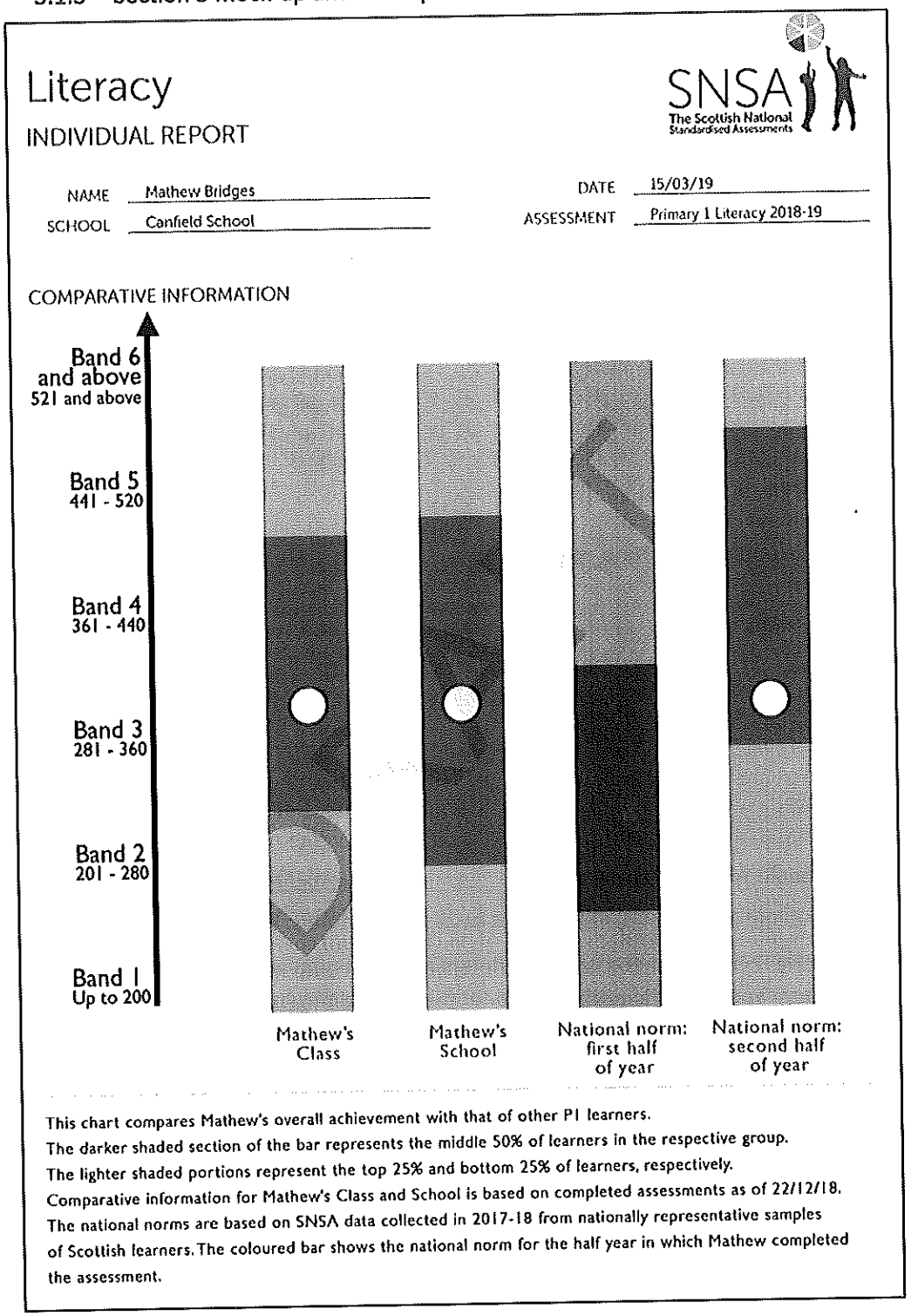
Descriptor	Difficulty	Result
Track a pronoun reference and interprets the meaning of 'but', in a single sentence , without audio support	Band 5	✘ Incorrect
Retrieve information from a 3-sentence narrative without audio support	Band 1	Not Attempted

220

This section displays the items presented to the child or young person, and his or her response on each of the items (Correct, Incorrect or Not attempted). Layout and content are as in AY 2017-18, except 'Band #' will replace the rating High, Medium or Low.[3] As in AY 2017-18, items will be arranged in descending order of difficulty within each organiser (not as shown here).

[3] This band label is an indicator of the item's difficulty. This learner, whose capacity is at Band 3, would typically be expected to succeed on Band 3 items twice as often as not; more often on items below Band 3, and less often on items above Band 3 (not shown here).

3.1.3 Section 3 mock-up and description

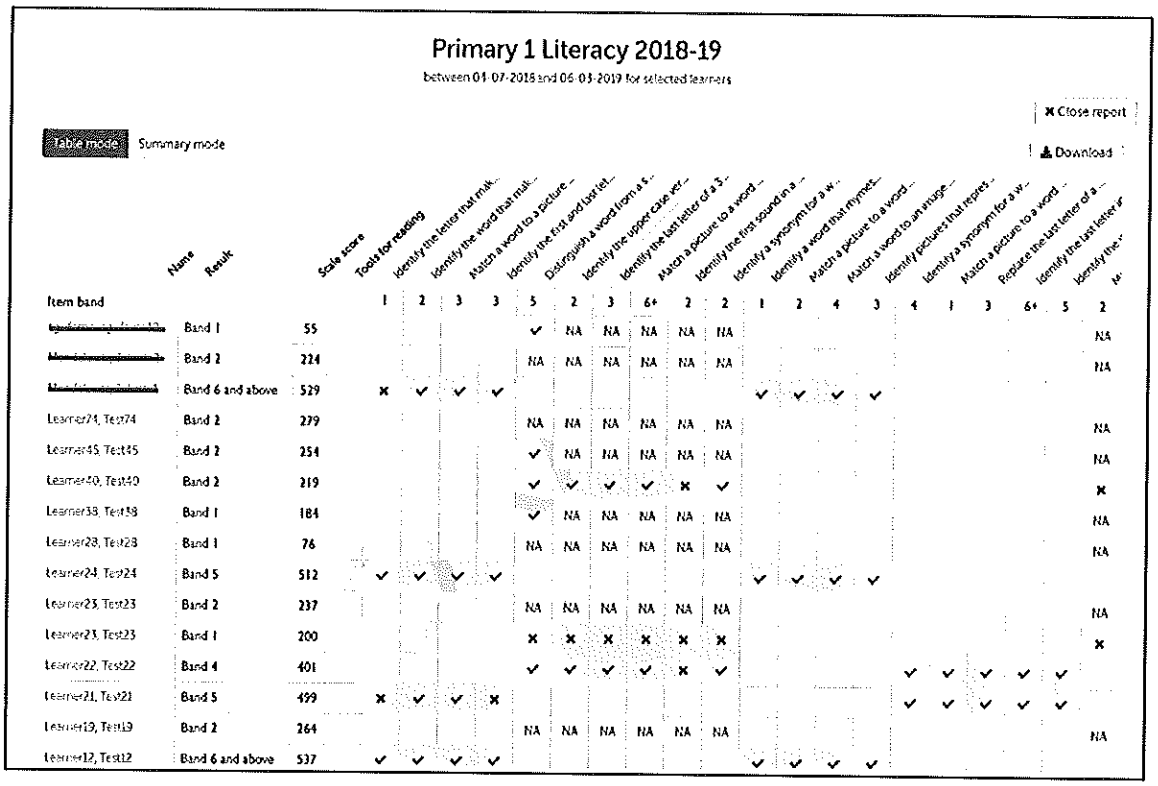


Section 3 of the individual report provides comparative information about the learner's capacity in relation to class, school and national norms for the two norming studies conducted with Scottish children and young people in AY 2017-18. Bars for a learner's class and school will only appear after a minimum of ten completions in that cohort, and will be labelled as in AY 2017-18 reports. Two bars showing national norms will be included in this section, one showing the norm for the first half year and one for the second. The individual's capacity will be shown on the half-year norm bar in which the learner completed the assessment. The other norm bar will be presented in greyscale. Text

below the graphic provides information on how to interpret it. (Please note that in the mock-up the dates are not accurate.)

3.2 Group diagnostic report

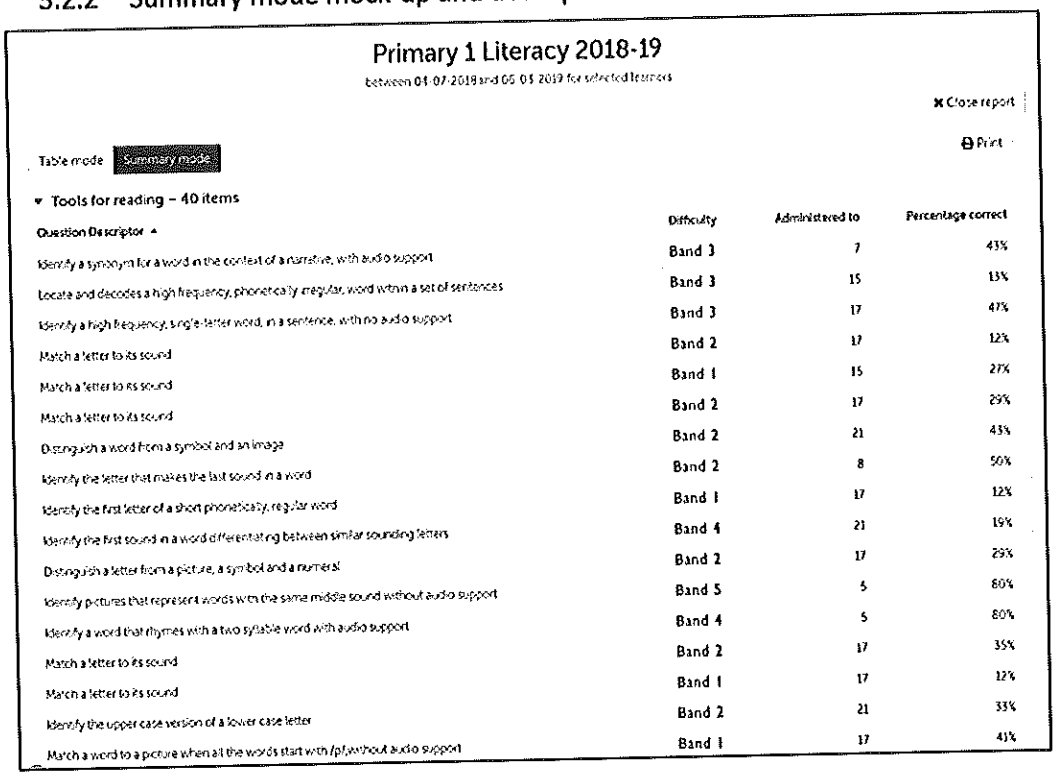
3.2.1 Table mode mock-up and description



In general, the layout and information provided will be as in AY 2017-18. The differences are:

- There will be a row beneath the item descriptors (diagonal text) which presents the band in which the item appears, as a quick indicator of the item's difficulty.
- The individual's capacity is shown as a band achieved(second column from left) and as a scale score (third column from left)
- In the pop-up (appearing when an item descriptor is hovered over), the item's band label will appear instead of High, Medium or Low.

3.2.2 Summary mode mock-up and description



Primary 1 Literacy 2018-19

between 04-07-2018 and 05-05-2019 for selected learners

Close report

Print

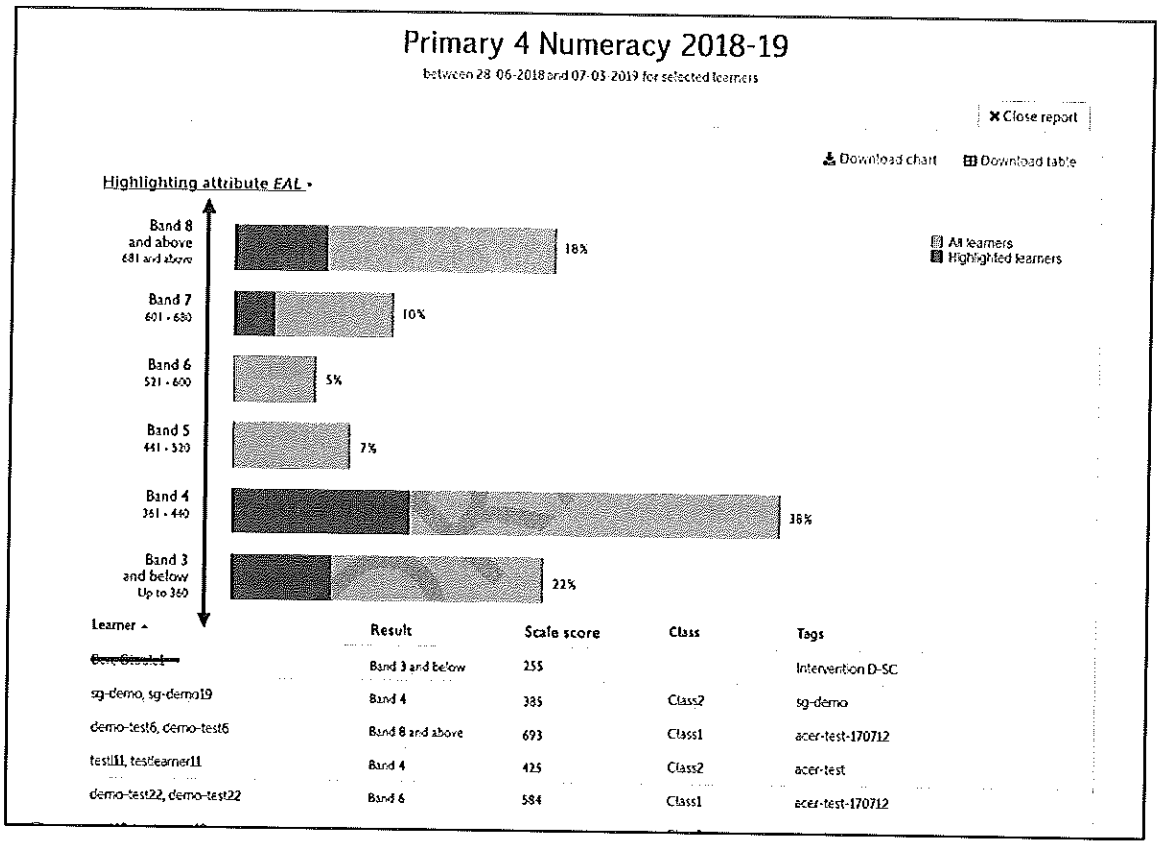
Table mode | Summary mode

▼ Tools for reading – 40 items

Question Descriptor ▲	Difficulty	Administered to	Percentage correct
Identify a synonym for a word in the context of a narrative, with audio support	Band 3	7	43%
Locate and decodes a high frequency, phonetically irregular, word within a set of sentences	Band 3	15	13%
Identify a high frequency, single-letter word, in a sentence, with no audio support	Band 3	17	47%
Match a letter to its sound	Band 2	17	12%
Match a letter to its sound	Band 1	15	27%
Match a letter to its sound	Band 2	17	29%
Distinguish a word from a symbol and an image	Band 2	21	43%
Identify the letter that makes the last sound in a word	Band 2	8	50%
Identify the first letter of a short phonetically regular word	Band 1	17	12%
Identify the first sound in a word differentiating between similar sounding letters	Band 4	21	19%
Distinguish a letter from a picture, a symbol and a numeral	Band 2	17	29%
Identify pictures that represent words with the same middle sound without audio support	Band 5	5	80%
Identify a word that rhymes with a two syllable word with audio support	Band 4	5	80%
Match a letter to its sound	Band 2	17	35%
Match a letter to its sound	Band 1	17	12%
Identify the upper case version of a lower case letter	Band 2	21	33%
Match a word to a picture when all the words start with /p/,without audio support	Band 1	17	41%

In general, the layout and information provided will be as in AY 2017-18. The main difference is that, instead of the High/Medium/Low rating of each item, its 'band' difficulty will be shown.

3.3 Group aggregate report: mock-up and description



In general, the layout and information provided in the group aggregate report will be as in AY 2017-18. The main difference is that, instead of three bars showing High, Medium and Low, the relevant bands for the target assessment will be shown (the same set of bands as in individual report).

In the mock-up presented here, for P4 Numeracy, the band labels should be 'Band 4 and below', 'Band 5', 'Band 6', 'Band 7', 'Band 8' and 'Band 9 and above' (not accurate in the mock-up), with an arrow at the top and bottom of the vertical line to the left. In this mock-up, EAL has been selected as a filter, so the darker pink shading shows the capacity of EAL learners in the selected group, with light pink showing other learners.

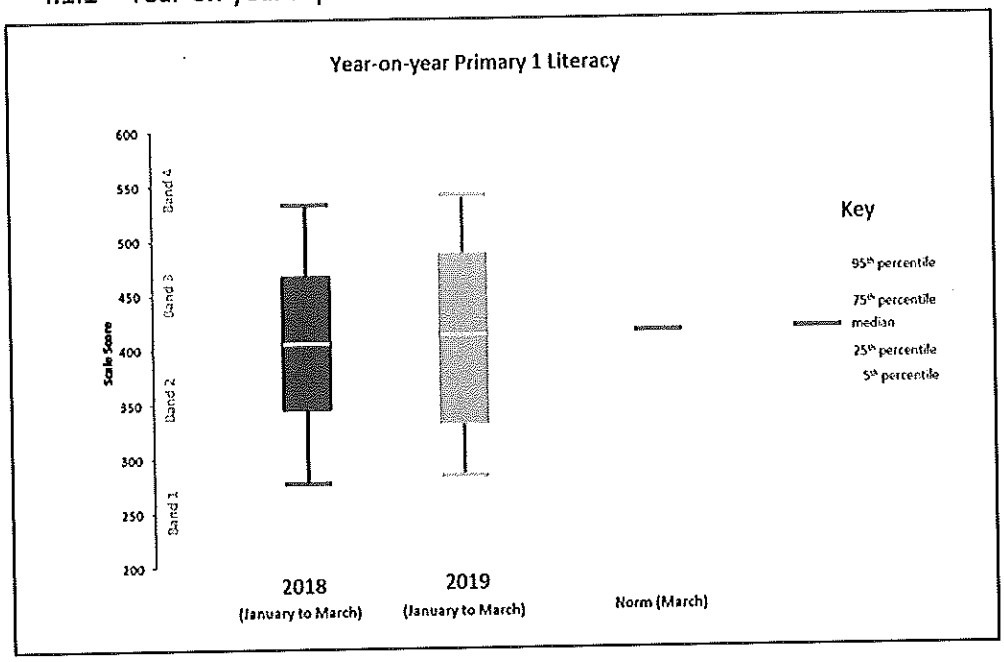
The tabular information below the graph lists all the children in the selected group – similar to the information provided on this report in AY 2018-19, but now with the learner's band and scale score presented, rather than High, Medium or Low.

4 School level report available later in AY 2018-19

4.1 Year-on-year report

This new report is planned for later in AY 2018-19 (date to be negotiated). Its purpose is to provide schools with information that allows them to compare the capacity of children in a given year group over time. Two different versions of this report are offered for comment: one for groups of 30 or more, and one for smaller groups.

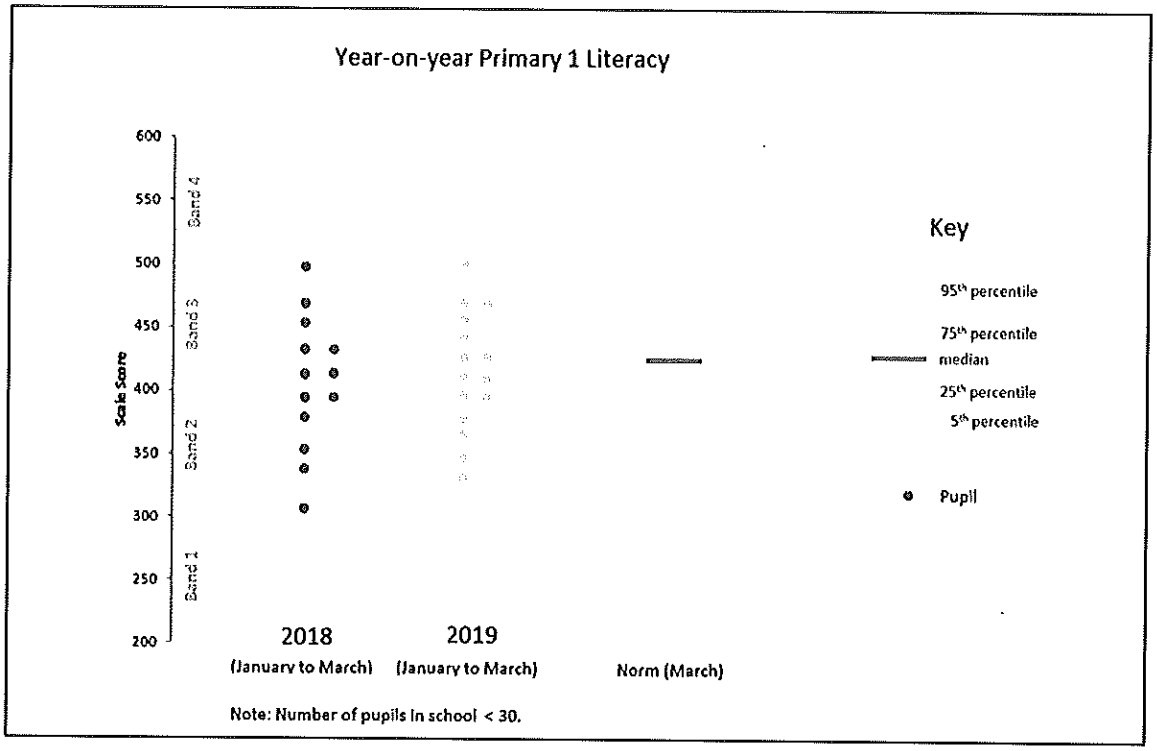
4.1.1 Year-on-year report for a cohort of 30 or more – mock-up and description



On the left hand side of this report, the bands and scale score ranges would be shown along a vertical arrow, as in the individual and group aggregate reports, with the bands and score range varying by year group. (In this mock-up, only a scale is shown.)

In the illustration, the range of capacities achieved by a Primary 1 cohort over two years is illustrated (for Scotland, this would be AY 2017-18 and AY 2018-19, rather than '2018' and '2019' as shown here). The November and March national norms could also be shown as a reference point (one norm shown here). The Key defines the box and whisker plots: here, the same definitions are used as in the LA Results Overview report (5th, 25th, median, 75th and 95th percentiles). As the SNSA goes into third and subsequent years of live assessments, more box and whisker plots could be added to this display.

4.1.2 Year-on-year report for a cohort of fewer than 30: mock up and description



285 This is similar to the previous year-on-year report. However, because of the small number in the cohort (fewer than 30), a box and whisker plot would easily give a distorted impression of the range of capacities demonstrated, so the learners' individual capacities are represented by dots instead.

Appendix

This appendix is intended to further explain how the equating to form a long scale is carried out. As outlined in the body of this paper, formation of a long scale depends on items behaving in a similar, parallel fashion with different year groups. To bridge the three-year gap between the target year groups of the SNSA, a long scale equating study was conducted, administering sets of SNSA items to year groups adjacent and adjacent-but-one to P1, P4, P7 and S3, as follows:

P1 items were administered to P2 and P3

P4 items were administered to P2, P3, P5 and P6

P7 items were administered to P5, P6, S1 and S2

S3 items were administered to S1 and S2.

In the analysis, the performance of items that have been administered to target year group, and the adjacent and adjacent-but-one cohorts, is compared. Those items that have performed similarly for the three year groups (target, adjacent, and adjacent-but-one) are used as the basis for equating.

To illustrate this, two examples from the SNSA analysis are provided. The first shows equating between P7 reading and S1&S2 reading, which indicated strong parallel performance of items across these year groups. The second shows equating between P1 numeracy and P2&3 numeracy, where the performance of only about half of the linking items indicated that they were suitable for linking. In each case, data from the two norming studies were used for P1 and P7 analysis, and data from the long scale equating study was used for the intervening year groups' analysis

Equating between P7 Reading and S1_S2 Reading

The following figures show the working for equating P7 Reading and S1_S2 Reading.

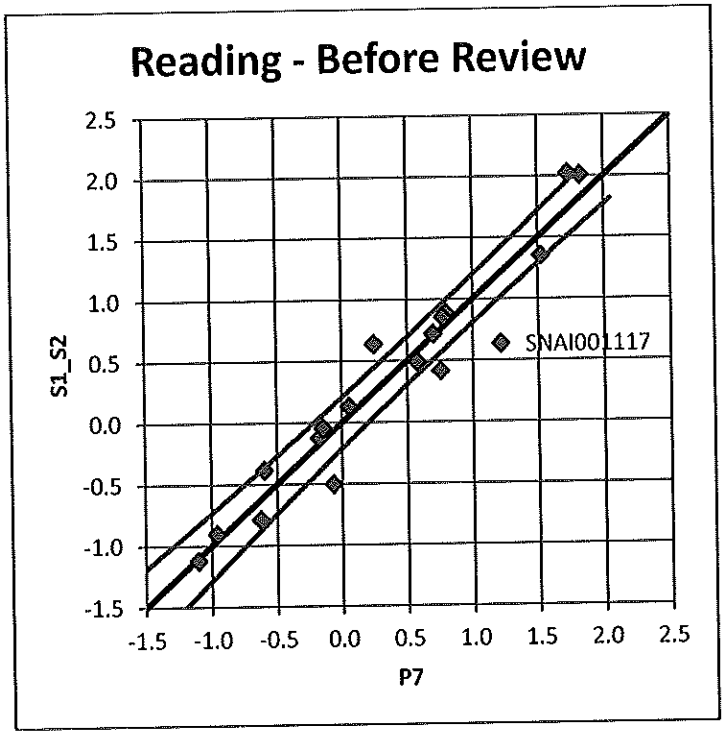


Figure 3 Comparison of P7 reading with S1&S2 reading, before review

Figure 3 shows how items in the P7 form compare with the same items in S1 and S2 forms. It can be seen that only one item is outside the chosen confidence interval.

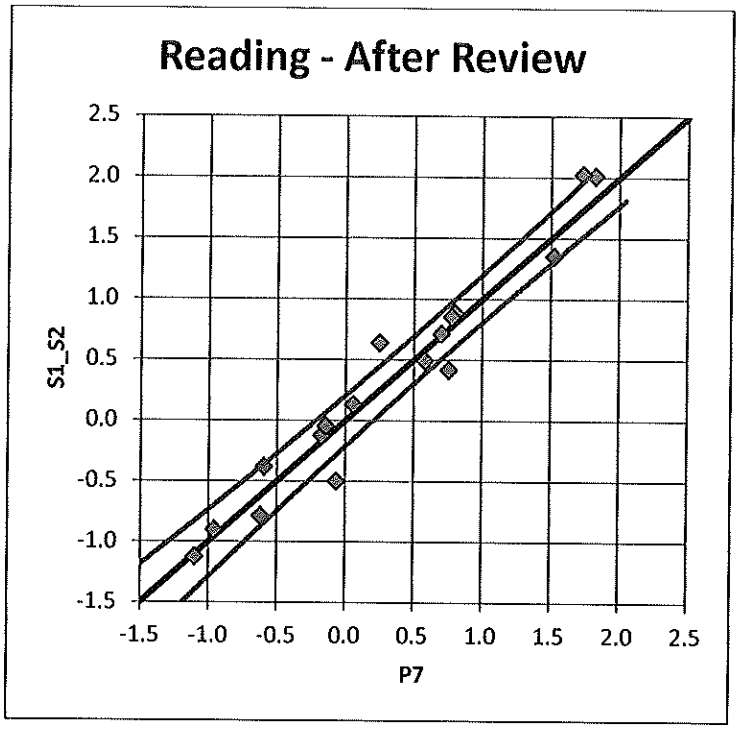


Figure 4 Comparison of P7 reading with S1&S2 reading, after review

Figure 4 shows how items in P7 form compare with the same items in S1 and S2 forms after the outlier was removed.

The statistics below show the difference in shift and SD ratio before and after the adjustment.

	Before Review				After review			
	CI (L)	Shift	CI (H)	SD Ratio	CI (L)	Shift	CI (H)	SD Ratio
Equating Shift in Logit	0.388	**0.501**	0.613	0.993	0.370	**0.469**	0.567	1.017

Equating between P1 Numeracy and P2_P3 Numeracy

35 The following figures show the working for equating P1 Numeracy with P2_P3 Numeracy.

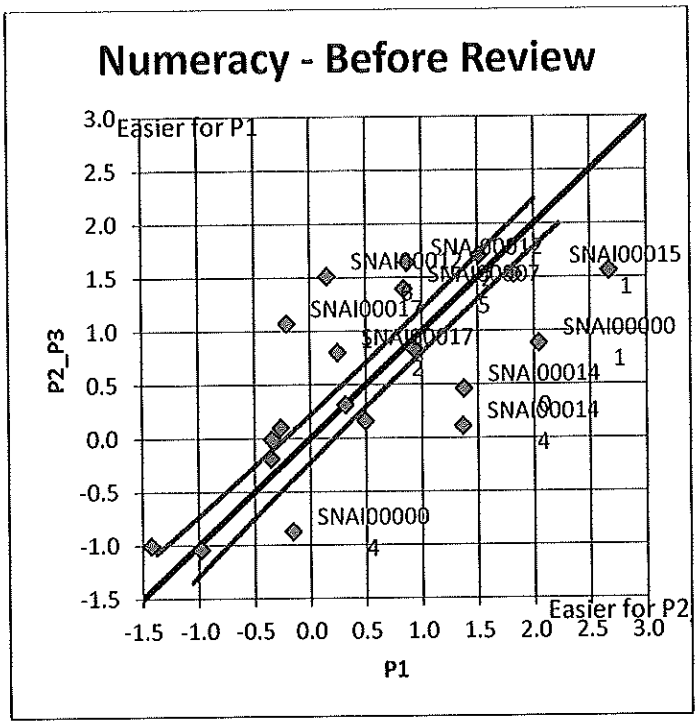


Figure 5 Comparison of P1 numeracy with P2&P3 numeracy, before review

Figure 5 shows how items in the P1 form compare with the same items in P2 and P3 forms. It can be seen that many items are outside the chosen confidence interval.

40

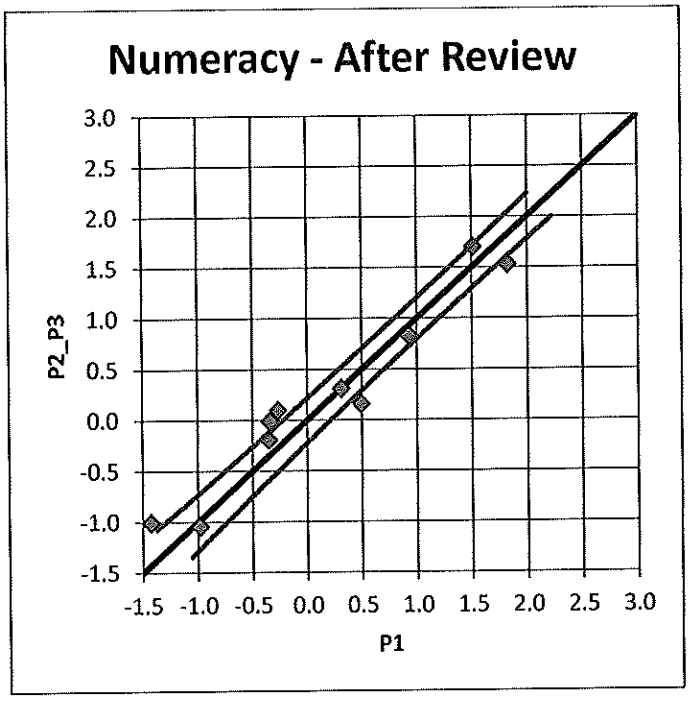


Figure 6 Comparison of P1 numeracy with P2&P3 numeracy, after review

Figure 6 shows how items in the P1 form compare with the same items in P2 and P3 forms after the
45 outliers were taken out.

The statistics below show the difference in shift and SD ratio before and after the adjustment.

	Before Review				After review			
	CI (L)	Shift	CI (H)	SD Ratio	CI (L)	Shift	CI (H)	SD Ratio
Equating Shift in Logit	0.999	1.331	1.662	0.851	1.098	1.265	1.431	0.881

November norming study - participation

DOCUMENT CONTROL
Version Control

Theme	Standardisation
PRODUCT TITLE	November norming study final participation summary
PRODUCT CODE	AR03
DOCUMENT No.	AR03-11

VERSION HISTORY

Date	Version	Comments	Redacted - exemption applies	Approved by	Date Approved
10/01/2018	0.1	November norming study final participation summary		BOT	11/01/2018

Final Numbers

NNS Assessment	NNS # assigned	NNS # sat	% of total # assigned
P1 Numeracy	800	688	86
P1 Literacy	800	696	87
P4 Numeracy	800	701	88
P4 Reading	800	705	88
P4 Writing	800	683	85
P7 Numeracy	800	680	85
P7 Reading	800	688	86
P7 Writing	800	669	84
S3 Numeracy	800	558	70
S3 Reading	800	561	70
S3 Writing	800	560	70
	8800	7189	82
Regular SNSA		318	4
Total	**8800**	**7507**	**85**

Summary of participation

LA	# of schools with sampled learners	# of schools who completed all of the norming assessments	# of schools who completed some of the norming assessments	# of schools who completed none of the norming assessments	% of schools per LA who did none of the norming studies	# of schools that did regular SNSA
Aberdeen city	60	39	13	8	13	2
Aberdeenshire	123	93	16	14	11	12
Angus	48	22	26	0	0	11
Argyll & Brute	49	32	4	13	27	1
Clackmannanshire	20	16	4	0	0	
Dumfries & Galloway	77	58	10	9	12	1
Dundee City	43	25	16	2	5	3
E Ayrshire	46	14	6	26	57	
E Dunbartonshire	43	25	10	8	19	4
E Lothian	37	23	11	3	8	1
E Renfrewshire	31	25	5	1	3	
City of Edinburgh	116	82	24	10	9	
Eilean Siar	15	8	2	5	33	
Falkirk	54	37	15	2	4	
Fife	133	100	27	6	5	4
Glasgow	174	98	48	28	16	
Highland	119	67	29	23	19	9
Inverclyde	26	17	8	1	4	
Midlothian	36	23	12	1	3	
Moray	38	23	7	8	21	3
N Ayrshire	53	44	7	2	4	
N Lanarkshire	141	95	31	15	11	
Orkney	10	10			0	
Perth & Kinross	61	47	10	4	7	2
Renfrewshire	60	42	13	5	8	3
S Ayrshire	40	33	6	1	3	
S Borders	55	42	10	3	5	
S Lanarkshire	128	106	16	6	5	
Shetland I	21	14	3	4	19	
Stirling	40	31	5	4	10	3
W Dunbartonshire	35	29	4	2	6	
W Lothian	75	13	42	20	27	27
Grand Total	2007	1333	440	234	12	86

Summary of non-response reasons

LA	SNSA	Absence	ASN	EAL	GME	IT Issue	Notice	Non-engagement	Move	Parent opt-out	School opt-out	Other	Grand Total
Aberdeen City	13	4						2					19
Aberdeenshire	68	5	3	1			3	3		1			84
Angus	35	4	1					1				1	42
Argyll & Bute	1	2			1	1					5	3	13
City of Edinburgh		10	8	3		1		2			2	1	27
Clackmannanshire		1											1
Dumfries & Galloway	1	3	5	1			5	1					16
Dundee	9	10	3			1						1	24
East Ayrshire		1		1		1	1					4	8
East Dunbartonshire	13					1							14
East Lothian	1	10						4					15
East Renfrewshire		2	5			1				1			9
Eilean Siar		2	1		3								6
Falkirk		4	4						1			4	13
Fife	11	5	1			2		1	2			4	26
Glasgow		16	15	2	5	1		1	1			7	48
Highland	22	5	4		3	4		2					40
Inverclyde		3						1				1	5
Midlothian		8				2							10
Moray	18												18
North Ayrshire		2	1										3
North Lanarkshire		14	5					2	1	2			24
Orkney													
Perth and Kinross	2	5	1		3	2			1				14
Renfrewshire	8	10	4					2		2			26
Scottish Borders		2				1							3
Shetland		2	1									1	4
South Ayrshire		3	2										5
South Lanarkshire		8	6			1						7	22
Stirling	18	2	1										21

West Dunbartonshire		2		1						1			4
West Lothian	92		2			2						1	97
Grand Total	**312**	**145**	**73**	**9**	**15**	**21**	**9**	**22**	**6**	**7**	**7**	**35**	**661**

SNSA March norming study analysis report

DOCUMENT CONTROL

Theme	Standardisation
PRODUCT TITLE	SNSA data analysis plan
PRODUCT CODE	AR03
DOCUMENT No.	AR03-14

VERSION HISTORY

Date	Version	Comments	Author	Approved by	Date Approved
04/09/18	0.1	First draft	Redacted - exemption applies		
10/09/18	0.2	Minor edits			

Contents

Introduction - Norm referencing

Norm referencing is an alignment of outcomes against distributions of achievement for population reference groups.

For SNSA, the immediate purpose of norm referencing is to provide an overall picture of Scottish learners' assessment results in numeracy, reading and writing skills against which assessment outcomes of individual children, and of subgroups such as classes or year groups in a school, or children in different SIMD groups, can be measured.

Sampling

With respect to establishing population norms, the ACER proposal envisaged the establishment of norms for each of the domains and year levels at a fixed time point during the school year. Furthermore our proposal was that the norming study be based on a probabilistic sample of learners. By definition, a probabilistic survey is one where each learner in the population has a *known, non-zero chance* of inclusion in the sample.

It was agreed to conduct two norming studies in the first year of SNSA implementation: one in November 2017, and one in March 2018. The chosen option for the norming studies had the following key features:

- Providing a collection of data from an adaptive online assessment administration in November 2017 and March 2018; and
- A systematic randomised selection of children and young people who complete the adaptive assessment in specified testing windows.

An equal probability sample of 800 learners across all schools in Scotland was selected for the norming study for each assessment domain (P1: Literacy and Numeracy, P4/P7/S3: Reading, Numeracy and Writing) and year group (8800 in total for November and 8800 in total for March 2018). Participants were sampled systematically from a list of all learners in Scotland using a 'random start, constant interval' selection method. The list was systematically organised (sorted) by important auxiliary variables to ensure implicit stratification of the sample by those variables. Local Authority (LA), SIMD, age and gender were agreed as stratification variables for the sampling of learners in the four year levels.

Data analyses for the March norming study

The March norming study was completed with a very satisfactory overall response rate (83%). The response data was representative of all LAs.

The psychometrics team conducted several initial analysis procedures based on the data collected from the study. The analyses included a review of the data input from the Horizon system as well as some additional data cleaning and processing to prepare the data sets for the psychometric review.

The norming study covered four different years (P1, P4, P7 and S3). A number of different domains was covered in the study: Numeracy, (early) literacy, reading, and writing (grammar and punctuation, and spelling). Appendix 1 shows all forms indicating the assessed domains at each year level, and the numbers of assigned and assessed learners.

The data summary report is attached as Appendix 1.

Following the initial database verification checks, the psychometrics team performed a scaling review of the March norming study data conducting an IRT-based analysis. The scaling review included the following steps:

1. *Scottish calibrations by year level and domain* (P1 L/N, P4 R/N/W, P7 R/N/W, S3 R/N/W): Item fit, item statistics and Item Characteristic Curves (ICC) were reviewed to assess the psychometric quality of assessment items with regard to the Scottish context.
2. *Comparisons of Scottish and pre-calibrated item parameters* (P1 L/N, P4 R/N/W, P7 R/N/W, S3 R/N/W):
   We reviewed the differences between standardised Scottish (norming study) and pre-calibrated item parameters per year level and domain and produced graphical displays (scatterplots with confidence interval lines). This enabled us to identify potential differences in item functioning between the underlying calibrations of the current delivery system and those based on the Scottish norming study.
3. *Review of differential item functioning* (P1 R/N, P4 R/N/W, P7 R/N/W, S3 R/N/W):
   The review of the occurrence of differential item functioning (DIF) by gender using an IRT model was conducted to investigate measurement equivalence of assessment items across gender groups in the Scottish context.

Item maps summarising the information from the calibrations are attached as Appendices 2-12.

In addition to the scaling review, we conducted a review of adaptive design. This involved a review of admissible paths (ABD, ABE, ABF, ACD, ACE, and ACF) to obtain percentages of learners in each path as well as mean scores and the distributions of assessment scores within each path in order to confirm the appropriateness of the adaptive design for the Scottish context.

The branching rules were also reviewed in order to check the extent to which they can be further optimised for the Scottish context.

Conclusions from the March norming study

This section reports on the conclusions from the data analyses of the March norming study.

Table 1 Assessments reliability results

Form	Reliability
P1 numeracy	0.836
P1 literacy	0.850
P4 numeracy	0.869
P4 reading	0.867
P4 writing	0.883
P7 numeracy	0.871
P7 reading	0.857
P7 writing	0.811
S3 numeracy	0.900
S3 reading	0.892
S3 writing	0.810

The first analysis that was conducted was to ensure the reliability of the assessments. The summary of the results presented in Table 1 shows that the reliability of all assessments forms was high (more

than 0.75). The reliability measure used in the analyses is based on item response modelling and indicates the degree to which items measure the same latent construct (i.e. the respective assessment domain). Reliabilities above 0.7 are typically considered as satisfactory and above 0.8 as excellent.

The second analysis conducted was focused on targeting – ensuring that the items used are appropriate for the learners' proficiency.

Appendices 2-12 show the Item Variable Maps for all forms which provides a mapping of Rasch item difficulty and learner capacity estimates. On the right hand side of the map, item difficulty parameters (indicating where a learner has a 0.5 probability of giving a correct response to an item) are shown on a logit scale according to the estimates of their difficulty, from the easiest (at the bottom of the map) to the most difficult (at the top of the map).

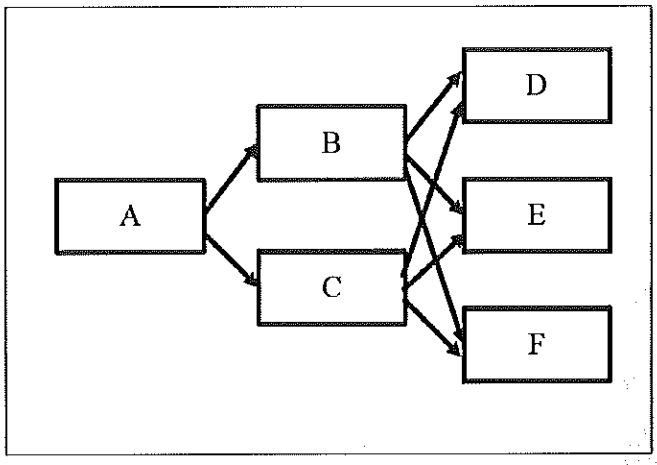
Analyses of the data from the learners in the norming study showed a wide distribution of achievement within the assessments. The analyses indicate that the items used in the study covered the range of capacity levels for learners within each domain and at each corresponding year level.

The Item Variable Maps show that while generally the distribution of learners' performances and the distribution of item difficulties are broadly at a similar level, there is an indication that there were somewhat more SNSA items matching the lower levels than the higher levels of learner capacity. The targeting differed somewhat across year levels and domains, with the closest match being between item difficulties and measured capacities at P1. Given that there was no prior information about targeting (except from the item trial in February 2017 based on convenience samples), achieving an ideal match between item difficulties and learner capacities has been a challenge since the inception of SNSA. We believe that the results generally indicate that each of the assessment forms is broadly at an appropriate level of difficulty for its corresponding group of learners, but that it is desirable to add more difficult items in the future, in order to make assessments more challenging and obtain more information about higher levels in each assessment domain.

Further analyses focused on the adaptive nature of the assessments, and reviewed the pathways within each assessment. Figure 1 illustrates the adaptive assessment design, and the branching process.

Figure 1 Adaptive assessment design

Adaptive assessment design



As illustrated in Figure 1, there are several possible paths that each learner can take, depending on how well they are performing in the assessment. Each cluster of items and each pathway is very carefully designed so that the whole assessment, for each learner, provides coverage of the required skills, as well as being matched to the demonstrated attainment of the learner as he or she progresses through the assessment.

As part of the analysis, the review percentages of learners in each path, as well as mean scores and the distributions of assessment scores within each path, were obtained, in order to confirm the appropriateness of the adaptive design for the Scottish context.

The results show that the adaptive system works very well in distinguishing between learners with more and less capacity: hardly any learners were found in the 'non-expected' paths from the more difficult middle cluster to the easiest one in the third position, or from the easier middle cluster to the most difficult one at the end. Thus, it can be concluded that the algorithm determining learners' capacity at the end of the starting cluster (A) leading them one to the most appropriate next testlet works well, and that there are hardly any cases indicating a misclassification of learners' capacity.

The higher percentages in the more difficult clusters (though at different levels across year levels and domains) provide support for the previous finding regarding the targeting of the assessment: learners are doing relatively well on the current assessments as the majority of learners are being presented with the more difficult clusters. Therefore the assessments would benefit from the increase in the overall level of difficulty within each cluster.

Next steps

Data from analysis will be used to inform two key areas:

1) It will be used in conjunction with data from the Long Scale Equating Study and November Norming Study to inform the construction of the long scale to be used in reporting for the 2018-19 school year onwards.
2) Data will be used to inform discussion between ACER and SG around the target difficulty of the assessments to be constructed for the next school year. Assessments for the next school

year, will be constructed from a pool of items comprising items from the current live assessments and items from the current in-test trial clusters which are deemed to be performing appropriately. ACER recognises that it may not be appropriate to move to the 'perfect' capacity model for the assessments for the next school year for several reasons:

a. noticeable change in difficulty between the first and second years of the assessments could be negatively received by teachers,
b. it is possible that the item pool may not contain a sufficient number of higher capacity items, given the views of previous question review panels in rejecting some items on the grounds they were too difficult for the target year group.

Appendices

Appendix 1 – Data summary report

Form	Non-participants	Participants	Total	% participation
P1 Numeracy	127	673	800	84%
P1 Literacy	132	668	800	84%
P4 Numeracy	130	670	800	84%
P4 Reading	119	681	800	85%
P4 Writing	113	687	800	86%
P7 Numeracy	102	698	800	87%
P7 Reading	140	660	800	83%
P7 Writing	136	664	800	83%
S3 Numeracy	154	646	800	81%
S3 Reading	163	637	800	80%
S3 Writing	177	623	800	78%
Total	1493	7307	8800	83%

Appendix 2 – Item Map for P1 Numeracy

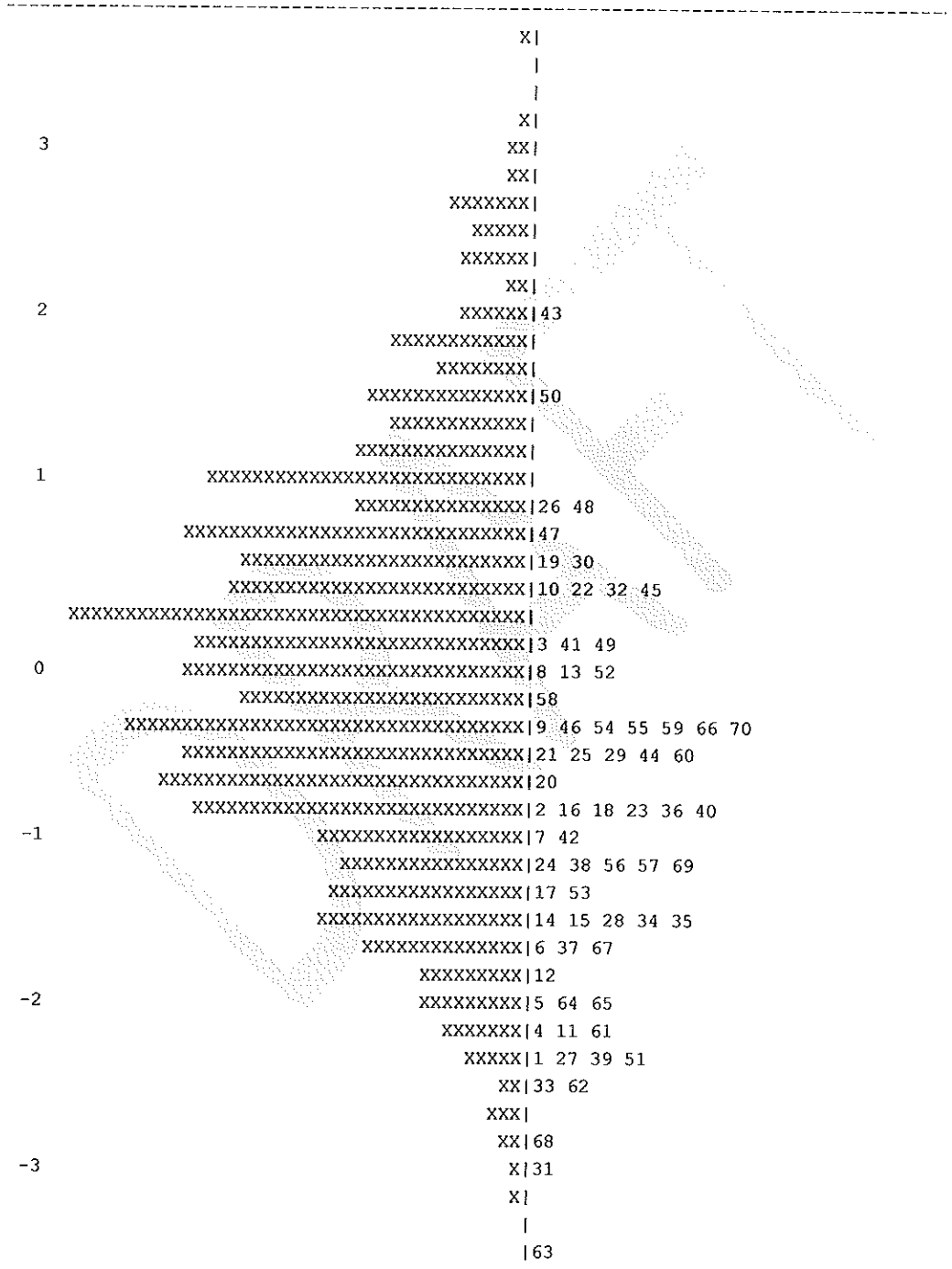
Each 'X' represents 0.9 cases

Appendix 3 – Item Map for P1 Literacy

Each 'X' represents 1.1 cases

Appendix 4 – Item Map for P4 Numeracy

Each 'X' represents 0.9 cases

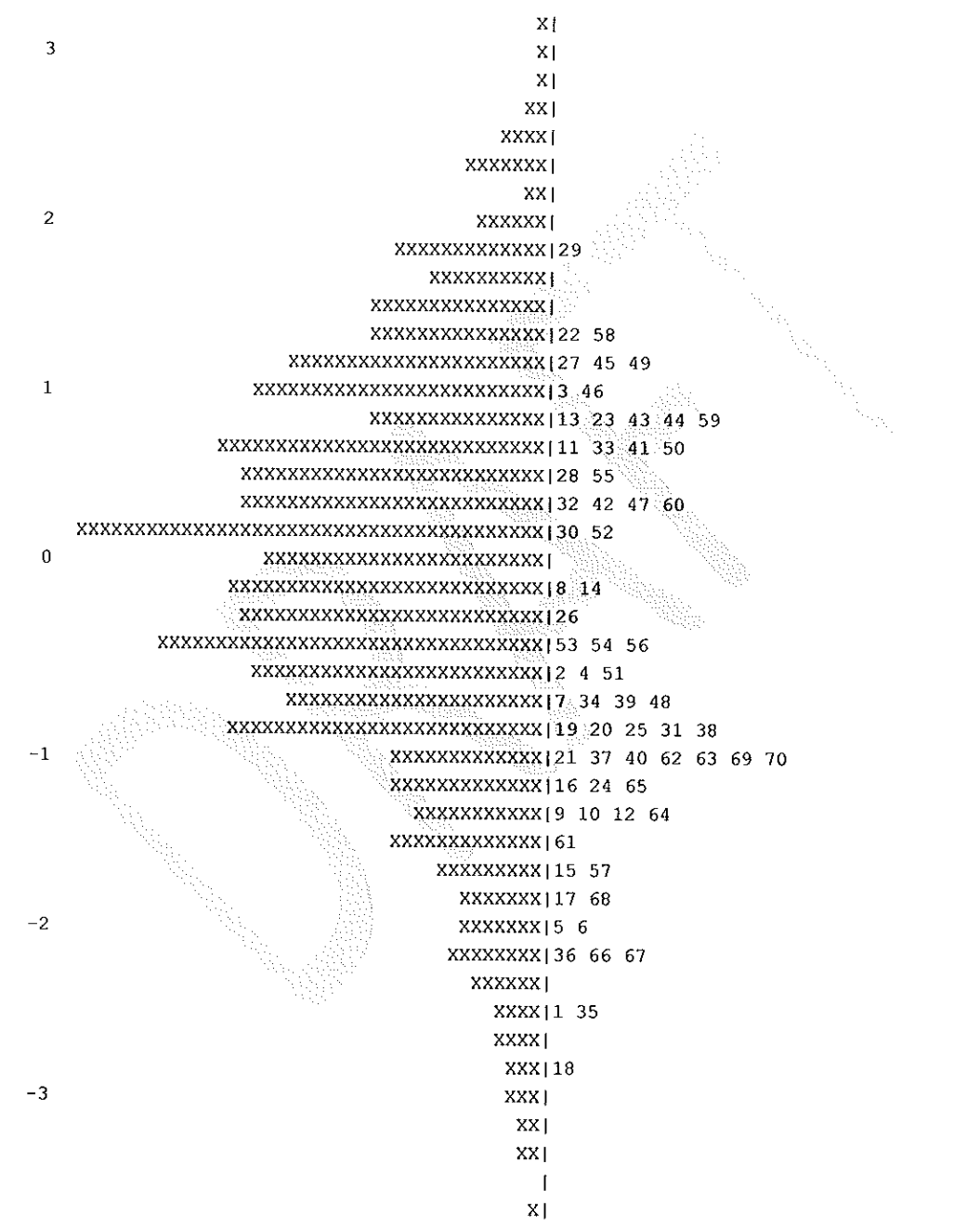


Appendix 5 – Item Map for P4 Reading

Each 'X' represents 0.9 cases

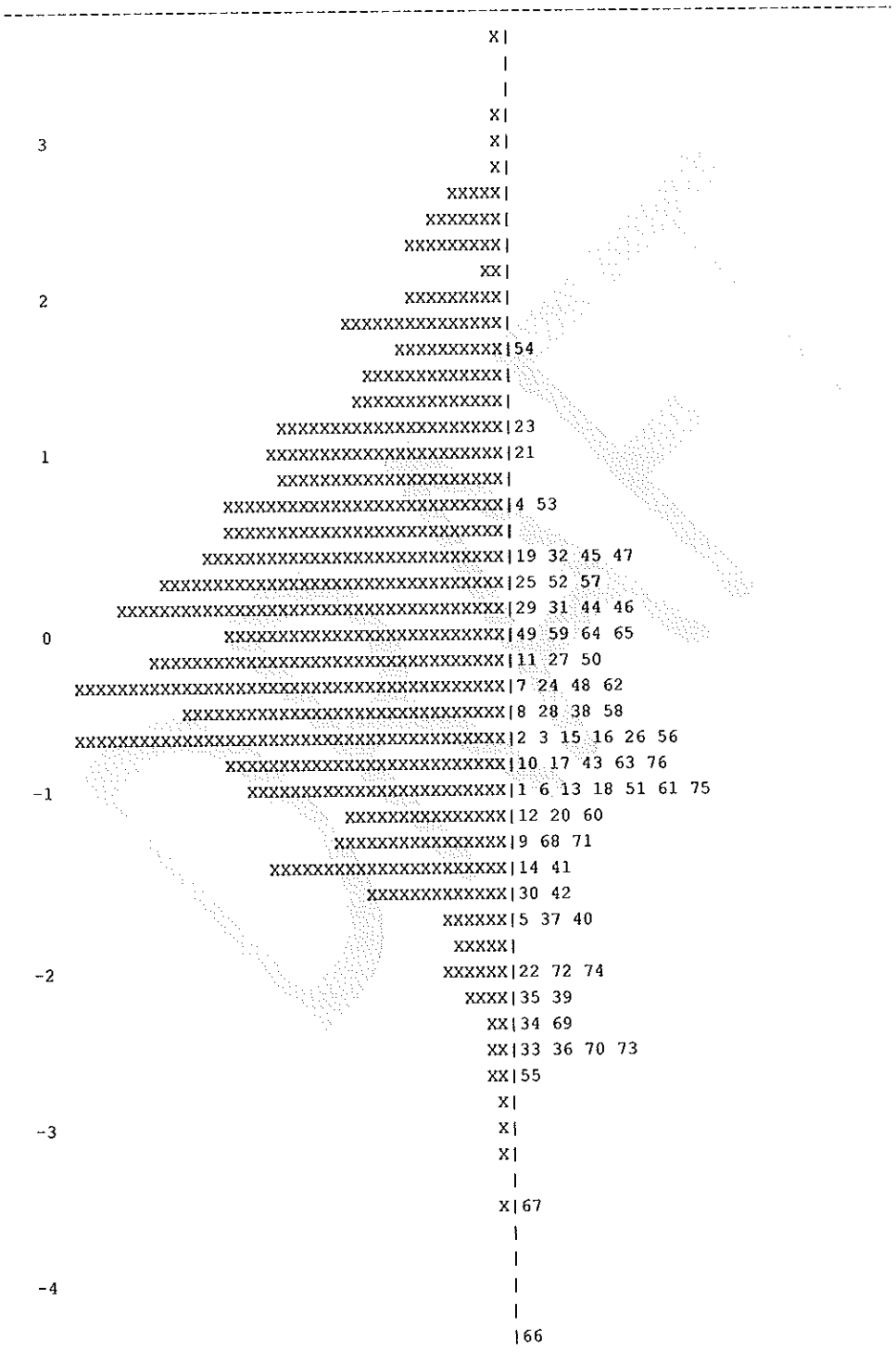
Appendix 6 – Item Map for P4 Writing

Each 'X' represents 1.0 cases



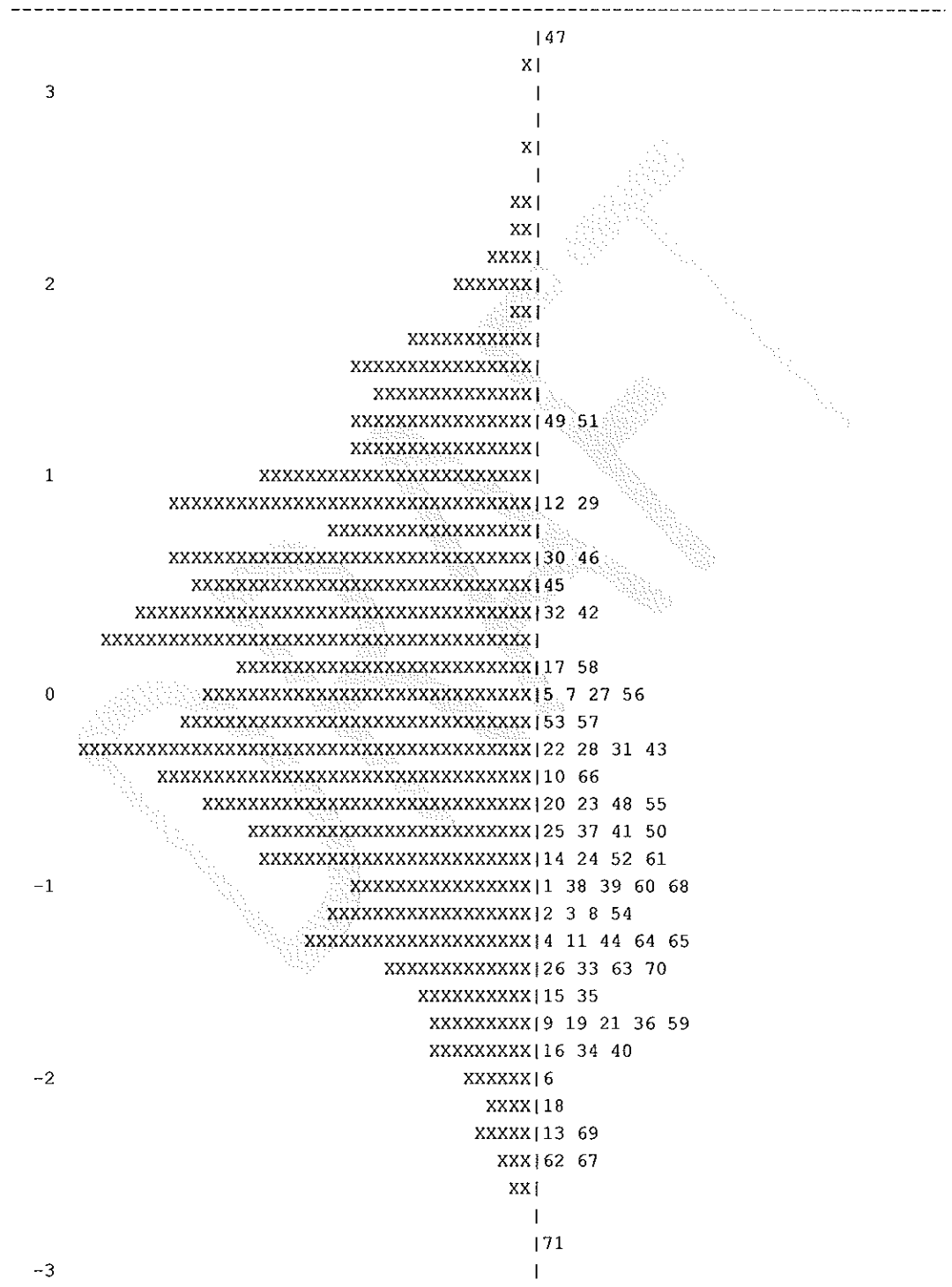
Appendix 7 – Item Map for P7 Numeracy

Each 'X' represents 0.9 cases



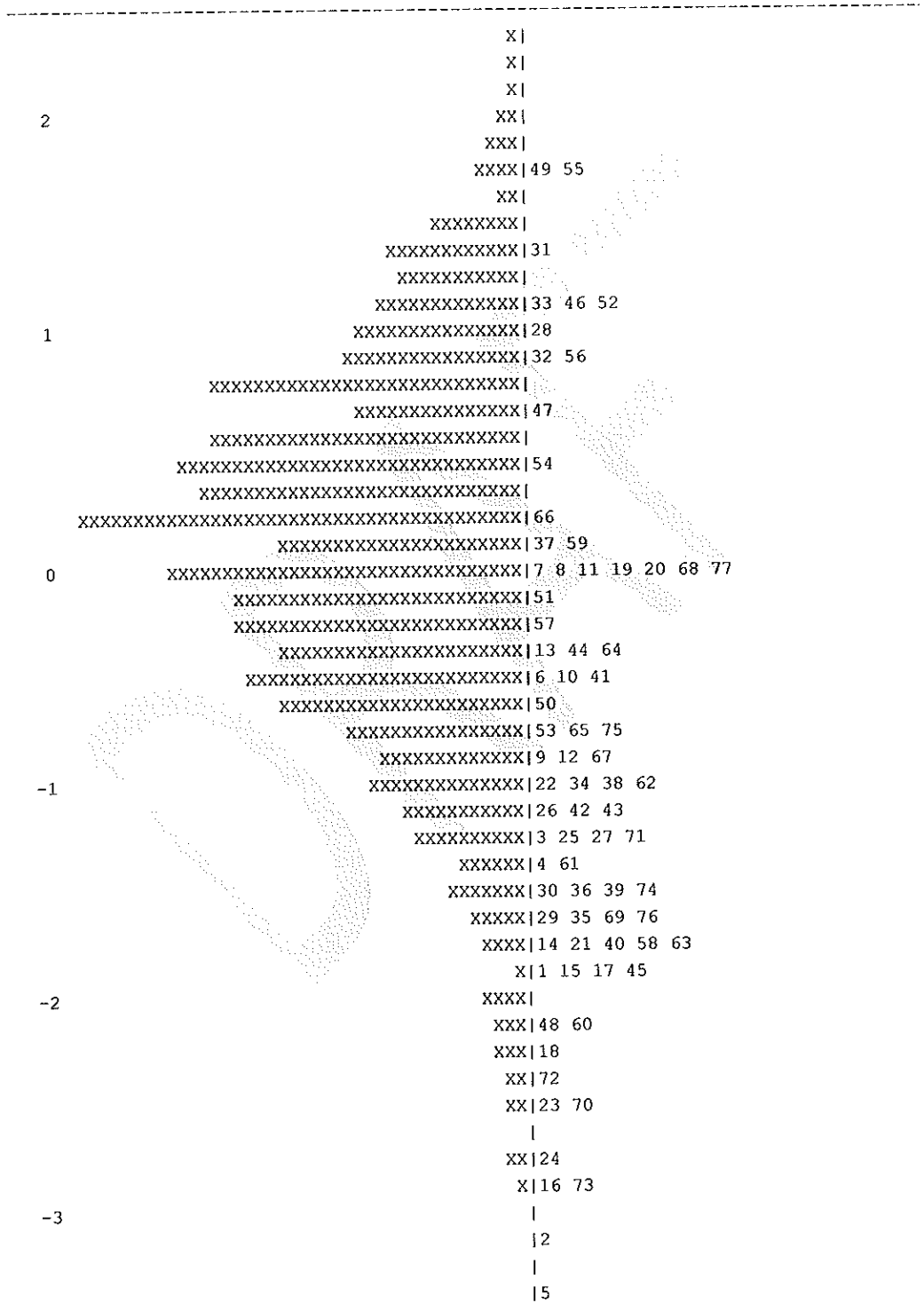
Appendix 8 – Item Map for P7 Reading

Each 'X' represents 0.8 cases



Appendix 9 – Item Map for P7 Writing

Each 'X' represents 1.0 cases



Appendix 10 – Item Map for S3 Numeracy

Each 'X' represents 1.0 cases

Appendix 11 – Item Map for S3 Reading

Each 'X' represents 1.0 cases

Appendix 12 – Item Map for S3 Writing

Each 'X' represents 0.9 cases