Analysis of Performance on P1 assessment items - 25/04/18

DOCUMENT CONTROL

PRODUCT TITLE	Analysis of Performance on P1 assessment items
PRODUCT CODE	AC04
DOCUMENT No.	AC04-07

VERSION HISTORY

Date	Version	Comments	Author	Approved by	Date Approved
25/04/2018	0.1	Initial Draft	[redacted]		
01/05/2018	0.2	Revised with feedback from [redacted]	[redacted]	[redacted]	09/05/18

Commented [MK(1]: section 38(1)(b) of FOISA (personal information)

Commented [MK(3]: section 38(1)(b) of FOISA (personal information)

Commented [MK(4]: section 38(1)(b) of FOISA (personal information)

Commented [MK(2]: section 38(1)(b) of FOISA (personal information)

1 Background

A review of item functioning is a standard part of the assessment development process and is conducted as standard on an annual basis. While similar analysis and interpretation is conducted for all year groups and subject areas, this paper is devoted to an analysis of Primary 1 item performance in order to feed into the current debate on the difficulty and appropriateness of the P1 assessments. It provides a summary of the behaviour of questions in the P1 assessments for the 2017/18 academic year.

The focus is particularly on aspects of the assessments about which concerns have been raised in various forums. Areas considered within this paper include:

1. The extent to which the difficulty of items, based upon the performance of Scottish children, corresponds to the initial low/medium/high classifications of question difficulty;
2. Review of drag and drop and matching items to determine whether functionality may have impacted upon performance on these items;
3. Review of items included in a cluster A to determine if performance on these items may have been unduly influenced by their position within the first 10 items in the assessment;
4. Identification of any general issues with the functioning of items;
5. For numeracy, review of items with charts and graphs to determine whether the level of audio support may have impacted on the performance of the items; and
6. For literacy, review of items which involved reading a text without audio support, to determine whether functionality may have impacted upon performance on these items.

The analysis used to inform the discussion in this paper is mainly based on data from live assessments completed up to 9 April 2018. There are some additional references to analysis from the November 2017 and March 2018 norming studies. The data referred to in the paper are from the analysis of live test completions unless specific reference is made to the norming studies.

2 Process for reviewing data for this paper

Findings in this paper are based upon review of initial analyses that are currently being generated by ACER's psychometricians as part of their work in constructing the long scale.

During the development of the SNSA forms for use in the 2017/18 academic year, items were classified as low, medium or high difficulty according to their original ACER scale score. For numeracy, items with an original ACER scale score in the range 90 and above were classified as high difficulty, items with an original ACER scale score in the range 71 to 89 were classified medium difficulty and items with an original ACER scale score in the range 70 and below were classified as low difficulty. For literacy, items with an original ACER scale score in the range 80 and above were classified as high difficulty, items with an original ACER scale score in the range 66 to 79 were classified medium difficulty and items with an original ACER scale score in the range 65 and below were classified as low difficulty.

For this paper the initial low, medium and high difficulty classifications were compared with difficulty estimates for the items based upon the performance of Scottish learners on data from the actual 2017/18 assessments. To do this the number of items assigned to each of the low, medium and high difficulty classifications in the current assessment were identified. The items were then categorised according to their difficulty based upon the performance of Scottish learners (called

henceforth SNSA live data set). Using this ordering the items were assigned a new, notional difficulty classification, such that the same number of low, medium and high classifications were assigned to match the original classifications. Items where the difficulty classification had changed between the original ACER scale score and the SNSA live data set were identified.

Additionally items were ranked by difficulty according to their original ACER scale score and the SNSA live data set. Items where there was a difference of eight or more places in rank ordering were considered for further investigation.

3 Numeracy

Items classified as 'high' difficulty tended to perform consistently according to their difficulty estimate from the SNSA live data set. There was greater variation in performance for items classified as low or medium, with five of the items classified as medium appearing among the 10 easiest items based upon the SNSA live data set.

3.1 Overall performance

Review of data from the November and March norming studies shows the difficulty range of the questions in the P1 numeracy assessment to be generally well matched to the capacity range of the learners. For the March data, performance shows an increased proportion of items at the lower end of the scale. As such the overall assessment could be regarded as easier than ideal.

Review of numeracy data according to the processes outlined in section 2 identifies curriculum areas of interest as follows.

- Scottish learners found some questions involving ordinal vocabulary more difficult than indicated by the original ACER scale score.
- Scottish learners tended to find items involving number recognition easier than indicated by the original ACER scale score.
- Scottish learners found some questions involving ordering and sequencing numbers more difficult than indicated by the original ACER scale score.
- Scottish learners tended to find items assessing understanding of positional vocabulary easier than indicated by the original ACER scale score.

There did not appear to be any consistent patterns in the performance of items reflecting other curriculum areas.

3.2 Performance on different item types

The assessment included two 'matching' items; these were of identical difficulty according to original ACER calibration, while one became easier for Scottish learners and the other more difficult in comparison with the original ACER calibration. These differences seem more likely to be due to curriculum content than item type, since the change in difficulty was inconsistent.

The assessment included 26 'drag and drop' items. Of these, according to the comparison of order of question difficulty, seven items became more difficult and one easier. It is noted that there is no consistent trend in the change in difficulty of drag and drop items, which again suggests that any changes are likely to be due to factors other than the ability to engage with the item type itself.

There is no way to know from these data how much support in recording answers was provided by class teachers, so it is possible that performance issues based on item type functionality may be masked. It should also be noted that this conclusion is based upon the assumption that children in other countries (whose responses were used to construct the ACER scale) did not have any issues with the drag and drop item format. Nevertheless, it is pleasing to note that there is no quantitative evidence that the nature of the item types is having an obvious adverse effect on Scottish learners' performance on these items.

3.3 Performance on items in the initial cluster

Sixteen items appeared in one or both of the 'A' clusters: that is, the initial cluster completed by the learners. Of these, four moved to a higher classification (i.e. changed from low to medium or medium to high) and one moved to a lower classification (i.e. changed from high to medium). A comparison of order of question difficulty, between the SNSA live data and the original ACER calibration, also shows four questions becoming more difficult. On this basis it is possible that an early position in the assessment is more likely to have a negative impact on performance than a later position in the assessment.

There are not enough questions from each curriculum area to be able to determine whether any specific areas were more likely to be subject to increased difficulty. Nevertheless, a recommendation in relation to this is to carefully consider the curriculum areas addressed by items in the initial cluster, to ensure that those curriculum areas are ones likely to be introduced in the first half of the school year, and thus will be familiar to the learners.

3.4 Issues with item functioning

Part of the analysis of items for each year of SNSA involves a review of the item statistics for each item to identify any items which are found to correlate poorly with the assessment as a whole or are so easy or difficult that almost all learners or no learners, respectively, are answering correctly. This type of review has been completed using the SNSA live data set and has identified the majority of items to be performing as expected.

A small number of items showed unusually low correlation between performance on the item and performance on the assessment as a whole (point biserial correlation). P1 numeracy items with this feature were:

[redacted]

Commented [MK(5]: section 33(1)(b)
6 sentences

Review of these items does not identify any obvious issues with the content of the items. The items relate to two specific curriculum areas: number sequences and visual comparison of measures. As such it is likely that these are curriculum areas that measure something different to the majority of items in the assessment and have a lower point biserial correlation for this reason rather than indicating specific issues with the items.

3.5 Performance on items with charts and graphs

There are six items in the assessment that address information handling. They are presented in difficulty order below, based upon item statistics from the SNSA live data set. Comparing difficulty with item content demonstrates all items to be performing as expected. The varying levels of audio support are not obviously affecting item difficulty, given that the easiest item is also the one with least

[redacted]

Commented [MK(6]: section 33(1)(b) – commercial interest
6 sentences

4 Literacy

4.1 Overall performance

Review of data from the November and March norming studies shows the difficulty range of the questions in the P1 literacy assessment to be generally well matched to the capacity range of the learners. For the March data, performance shows an increased proportion of items at the lower end of the scale and few items matching to the higher end of the scale. As such the overall assessment could be regarded as easier than ideal.

Review of P1 literacy data according to the processes outlined in section 2 identifies the following curriculum area of interest.

- Scottish learners found some questions involving matching words and images more difficult than indicated by the original ACER scale score.

There did not appear to be any consistent patterns in the performance of other curriculum areas.

4.2 Performance on different item types

The assessment included one 'matching' item; comparing performance between the original ACER calibration and the Scottish live data set indicated performance to be similar.

The assessment included 23 'drag and drop' items. Of these, according to the comparison of order of question difficulty, seven items became more difficult and four easier. It is noted that there is no consistent trend in the change in difficulty of drag and drop items, which suggests that any changes are likely to be due to factors other than the ability to engage with the item type itself.

As mentioned for numeracy, there is no way to know from these data how much support in recording answers was provided by class teachers, so it is possible that performance issues based on item type functionality may be masked. It should also be noted that this conclusion is based upon the assumption that children in other countries did not have any issues with the drag and drop item format. Nevertheless, it is pleasing to note that there is no quantitative evidence that the nature of the item types is having an obvious adverse effect on Scottish learners' performance on these items.

The other two items types included in this assessment were 'hot spots', where the learner clicks on a card to select their answer; and 'wordy' where the learner clicks on a word to select their answer. For both of these items types three items were found to be easier and none more difficult.

4.3 Performance on items in the initial cluster

Twenty items appeared across the different 'A' clusters: that is, the initial cluster completed by the learners. Of these, three moved to a higher classification (i.e. changed from low to medium or medium to high) and one moved to a lower classification (i.e. changed from medium to low). A comparison of order of question difficulty, between the SNSA live data and the original ACER calibration, shows two questions becoming more difficult and three becoming easier. There are not enough questions from each curriculum area to be able to determine whether any specific areas were more likely to be subject to a change in difficulty.

On this basis there is no evidence to suggest that an early position in the literacy assessment is more likely to have a negative impact on performance than a later position in the assessment.

4.4 Issues with item functioning

Part of the analysis of items for each year of SNSA involves a review of the item statistics for each item to identify any items which are found to correlate poorly with the assessment as a whole or are so easy or difficult that almost all learners or no learners, respectively, are answering correctly. This type of review has been completed using the SNSA live data set and has identified the majority of items to be performing as expected.

A small number of items showed unusually low correlation between performance on the item and performance on the assessment as a whole (point biserial correlation). Items with this feature were:

[redacted]

Commented [MK(7]: section 33(1)(b) – commercial interest
3 sentences

Review of these items does not identify any obvious issues with the content of the items.

Additionally there is one item where performance in the live assessment is anomalous with performance on the same item in the norming study data. This item will be subject to further investigation and will be reported on separately.

- **[redacted]**

Commented [MK(8]: section 33(1)(b) – commercial interest
1 sentence

4.5 Performance on reading texts with and without audio support

There are 27 reading comprehension items within the assessment. These comprise seven items associated with five single sentence / short texts and 20 items associated with four longer texts (three stories and an information text).

Of the sentence items there is one with audio support; the difficulty of this item is consistent across the SNSA live data set and the original ACER scale used to construct the assessments. Of the six sentence items without audio support, one was found easier in the SNSA data set than the ACER calibration, in terms of both the order of item difficulties and a change in difficulty classification. There were two items that became more difficult, and a third one where the difficulty classification changed. The items that became more difficult were all presented in the initial cluster while the one that became easier was presented in a second phase cluster.

Of the four reading texts, two included audio support and two did not. Two of the texts appeared in each of the second phase clusters presented to the learners (cluster B and cluster C) and the other two appeared in two of the third phase clusters presented to the learners (cluster D and cluster E).

- **[redacted]** appeared in the lower difficulty second phase cluster (C), and had audio support. None of the items in this text were identified as changing difficulty. The items were shown to be of a medium/low difficulty in the SNSA live data set, which is as expected.
- **[redacted]** appeared in the higher difficulty second phase cluster (B), and did not have audio support. Two of the items associated with this text were found more difficult than expected according to the original ACER scale score. The items were shown to be of a high difficulty in the SNSA live data set, which is consistent with the text's location in this cluster. One of these items was found to be the most difficult in the assessment.
- **[redacted]** appeared in the medium difficulty third phase cluster (E) and had audio support. One of the items was found more difficult than expected according to a comparison of the original ACER calibration and the SNSA live data set. Another item became easier, moving from a high to medium difficulty classification. The items were generally shown to be of a medium difficulty in the SNSA live data set, which is consistent with the position of the text in this cluster.
- **[redacted]** appeared in the highest difficulty third phase cluster (D) and did not have audio support. Two of the items associated with this text was found more difficult than expected according to the original ACER scale score. The items were shown to be of a high difficulty in the SNSA live data set, which reflects the text's position in the highest difficulty cluster.

Commented [MK(9]: section 33(1)(b) – commercial interest 1 sentence

Commented [MK(10]: section 33(1)(b) – commercial interest 1 sentence

Commented [MK(11]: section 33(1)(b) – commercial interest 1 sentence

Commented [MK(12]: section 33(1)(b) – commercial interest 1 sentence

It is noted that there has been criticism from some schools around the lack of audio support for some of the reading comprehension items in the assessments, and that this had not been expected given that all of the examples in the practice assessment did have audio support. Inspection of the data, however, does not identify a consistent pattern in changes in difficulty for items without audio support; rather, the SNSA live data set shows items for each text to be performing in the expected difficulty range, which suggests that the lack of audio for these items did not have a confounding impact for learners working in the targetted capacity range.

5 Conclusions

This purpose of this paper has been to analyse Primary 1 item performance in order to feed into the current debate on the difficulty and appropriateness of the P1 assessments, focusing especially on features of the assessments about which concerns have been raised.

Aspects of the P1 assessments investigated in this paper generally do not provide psychometric evidence in support of these specific areas of concern with the assessments.

The review of item functioning for the numeracy and literacy assessments has identified a small number of items which will merit further investigation when deciding upon the items to be available for selection for the 2018/19 assessments. This is to be expected given that the items were selected based upon existing ACER calibrated data from countries other than Scotland. This situation may also be witnessed in future years, since it is not uncommon for the performance of items to change over time as emphasis on the teaching of different curriculum areas varies over time. Another reason for changes in item performance is that items may become 'over-exposed': that is, the content of the item becomes widely known so learners are giving the expected answer for the item rather than actually interacting with the item to work out the answer.