Technical Information

PT SCIENCE

Trials

The questions in *Progress Test in Science* (*PTS*) were developed by science subject experts for a question bank acquired by GL Assessment. The National Foundation for Educational Research (NFER) then selected and adapted the material for use in *PTS* and also created some new test questions. The questions cover the different content domains of the science curriculum and in addition address three reporting areas.

The purpose of the trial was to gain evidence on a variety of aspects of the materials:

- to confirm the suitability and manageability of the items;
- to gain information about the functioning of the items with which to inform item selection;
- to provide data to enable age standardisation and other performance measures.

The trials took place in March 2015 and for each test level two forms (versions A and B) of both paper and digital formats were trialled. To ensure equivalency, the item formats in both forms were limited to true/false and multiple-choice (with pupils selecting either just one correct response or selecting two correct responses). The tests were not timed but 60 minutes was recommended to class teachers as the maximum time allocation.

Stratified random samples were drawn to trial the tests. These samples covered England, Wales, Scotland and Northern Ireland. Approximately half of the pupils that trialled the tests were from the target year and the remaining pupils were drawn from the year above. This was done to ensure that a sufficient proportion of the pupils had been taught the entire curriculum for the target year. Because the trial took place in the middle of the school year, only around half of the target year science curriculum would have been covered by the pupils in the target year.

Schools were asked to administer one test booklet (either version A or B) for each level they were trialling. The numbers of students taking part in the trials were as follows:

Test level	Number of students				
	Paper A	Paper B	Digital A	Digital B	Total
PTS8	349	343	302	243	1237
PTS9	376	396	260	336	1368
PTS10	377	403	227	234	1241
PTS11	304	367	253	289	1213
PTS11T	171	152	117	142	582
PTS13	628	587	183	183	1581
PTS14	566	502	164	150	1382

Standardisation

The data from the trials were analysed to provide information on the difficulty level of each question, its ability to discriminate between high and low scorers, and the extent to which it proved equally difficult for both genders, once each sex's general level of performance was taken into account. This information was then used to select questions for the final standardisation version of the paper and digital test. The items for the final test have been selected to achieve the best balance between overall test difficulty, proportions of the content areas and performance of individual items. The paper and digital versions of the tests have the same content. Further analysis showed that most of the questions showed no differential performance between paper and digital. Therefore, the paper and digital trial data was combined for the final standardisation.

The standardisation of all levels (except *PTS*11T) was based on the trial data. Within each level, the A and B forms from the trials had some common questions. With the common questions it was possible to combine the results of both versions and use a statistical model (Item Response Theory) to estimate scores based on the questions selected for the final standardisation version.

*PTS*11T is developed for use at the beginning of the school year in September. The A and B forms for this level were trialled in March 2015. A separate standardisation exercise was carried out in September 2015 based on the final version of the test. In the standardisation study 5,719 students took the paper version and 1,976 students took the digital version of *PTS*11T.

Test reliability

The reliability of a test is a measure of the consistency of a student's test scores over repeated testing, assuming conditions remain the same – that is, there is no fatigue, learning effect, or lack of motivation. Tests with poor reliability might result in very different scores for a student across two test administrations.

The reliability of the test was estimated using the Cronbach's Alpha formula which produces values ranging from 0 to 1. Values above 0.80 are considered to be very good. The reliability values for the *PTS* levels are given in the table below. They all show that the tests are very reliable. There were no significant differences between the reliabilities of paper and digital versions.

Test level	Reliability
PTS8	0.80
PTS9	0.83
PTS10	0.86
PTS11	0.88
PTS11T	0.90
PTS13	0.88
PTS14	0.86

For interpreting the score of an individual student, the standard error of measurement (SEM) is a more useful statistic than a reliability coefficient. It indicates how large, on average, the fluctuations in standard scores may be and indicates the 68% chance or confidence band. However, most tests show the 90% chance or confidence bands.

For example, the SEM for *PTS*10 is 5.6 for the UK. For an average-performing student with a *PTS*10 Standard Age Score (SAS) of 100, there is a 90% chance that the student's true SAS will be in the range +/- 9.2, i.e. between 91 and 109.

Test level	SEM	90% SAS confidence band (+/-)
PTS8	6.7	11.0
PTS9	6.2	10.2
PTS10	5.6	9.2
PTS11	5.2	8.5
PTS11T	4.7	7.8
PTS13	5.2	8.5
PTS14	5.6	9.2

Gender differences

The tests have been age standardised to a national mean of 100 and standard deviation of 15. There were approximately similar numbers of males and females in the standardisation samples. The table below shows the mean SAS score differences between males and females. Differences of more than 3 SAS points can be considered to be significant.

Test level	mean SAS differences
PTS8	2.1
PTS9	1.8
PTS10	-1.1
PTS11	-1.9
PTS11T	2.1
PTS13	-1.1
PTS14	0.4

Note: positive scores indicate females scored higher than males and negative scores indicate females scored lower than males.