GL assessment
the measure of potential

A SHORT GUIDE TO STANDARDISED TESTS

Published by GL Assessment Limited

389 Chiswick High Road, 9th Floor East, London W4 4AL

www.gl-assessment.co.uk

GL Assessment is part of the GL Education Group

Printed in Great Britain

A short guide to standardised tests

Introduction

A large number of schools across the UK choose to use standardised tests to augment their internal assessment regime. Standardised tests are used alongside teacher assessment and internally set, curriculum-linked tests, as well as data from national tests such as the English and maths SATs in England.

Some standardised tests are linked to key areas of the curriculum[1] and will test knowledge acquired against the criteria of a subject's programme of study. The majority of tests, however, are used either to test the underlying skills needed to make progress in learning, such as reading[2] or the abilities which support intellectual development, such as reasoning[3].

Any standardised test will go through rigorous development and take between two and four years to complete. The test structure has to be modelled, a large amount of test content must be developed and trialled with students in schools and then refined through a statistical process to produce the final tests. These are standardised on a very large, representative sample of students usually across the UK. The final normative data are produced from this final stage at which point supporting material for the teacher is developed. This often includes comprehensive reports offering further analysis of the test results.

Standardised tests are developed in a very structured way and are designed to test discrete skills and abilities in formats that have become established as fair and relevant.

Using standardised tests

Benefits of standardised tests

There are a number of key benefits of using well-developed standardised tests as part of a school's assessment regime. The key benefits are:

- The information such tests provide is quantifiable, that is it is in the form of scores and levels such as the Standard Age Score and National Curriculum levels provided by GL Assessment's reading, English and maths tests.
- Test results put an individual or group of students in the context of their peer group nationally and in some cases internationally. It is very important that schools know how their students' performance compares with other children of the same age across the UK and, increasingly, internationally[4].
- Using standardised tests over time allows progress to be tracked in an efficient, objective way. This applies to testing an individual at regular intervals or testing a particular year group to find out how, from group to group, performance may change.

1 GL Assessment's *Progress in Maths* series
2 GL Assessment's *New Group Reading Test*
3 GL Assessment's *Cognitive Abilities Test:* 4th Edition
4 http://www.oecd.org/pisa

- Standardised tests offer a reliable way of benchmarking a student's performance before intervention and an equally reliable way of assessing the impact of that intervention at a later date.
- The information from standardised tests may be interpreted and applied to an individual or group to improve teaching and learning.

Limitations of standardised tests

Standardised tests are only part of a complex system of assessment and should never be the only piece of information used to make decisions about performance.

- Any test will reflect a student's performance at a particular point in time and this may well be affected by non-cognitive factors such as fatigue and illness.
- Some students with Special Educational Needs may be unable to access the test.
- Some students with very high ability will reach the 'ceiling' of the test so the information from the test is not helpful to the teacher.
- Some tests do not give information that can feed meaningfully into changes to the curriculum or methods of teaching so practically, they may not be very useful.
- Test conditions may disadvantage some students who, under other circumstances, may perform at a higher level.

These limitations need to be considered when choosing to use a test and then when deciding how to use the test results. This means that testing must be part of a whole process which looks at the individual student as a person and as a learner through the quality of work in the classroom and through his teacher's professional judgements about performance and progress.

The role of standardised tests in evaluating intervention

Using a standardised test that offers equivalent forms is an ideal method of objectively measuring the impact of an intervention.

It is imperative that the test is age-appropriate and measures the skills which the intervention is seeking to improve.

Because the standard age score (SAS) is a recognised benchmark against a national sample of students of the same age, comparing this score before and after the intervention has taken place will allow a judgement to be made about progress achieved. A score of 100 at each test point shows that the level of attainment has been maintained; a lower score at the second point of testing does not necessarily mean that progress has not been made but a score that is lower by 8 standard score points or more is usually significant. Conversely, a gain in score of 8 standard score points or more may usually be considered as significant progress, however, the further the initial score is from the average, the greater the change has to be, to be considered significant. For example, a SAS of 78 from a SAS of 70 would not represent significant progress;

whereas a SAS of 108 from an SAS of 100 would. Students with very high scores will not be able to demonstrate progress as they will most likely reach the 'ceiling' of the test – that is they cannot show their improvement on this particular test.

Some students' scores will be lower at the second point of testing but not significantly so. In such cases, it is important to check the confidence bands (See page 7) around both sets of score: if these overlap then the scores are not significantly different.

A test is considered reliable if we get the same or similar result repeatedly.

Reliability refers to the consistency of a measure. It is impossible to calculate reliability exactly, but it can be estimated in a number of different ways. One way is to re-test a group of students and see how well the scores correlate between the two occasions. Reliability values go from 0 to 1, and the higher value is better. Higher reliability values lead to smaller confidence bands. For most of our tests the reliability values are in the region of 0.9 which is high. We usually refer to 90% confidence band which practically means that on 9 out of 10 occasions the true value of the score is within the score band.

Group versus individual testing

Group testing is time-efficient but may not be appropriate in all situations. If a student has very weak skills, the information from a group test may be irrelevant as the test is designed to include the majority of students but not always those with very low or exceptionally high levels of ability. In such cases, an individual diagnostic test should be used. However much as it may be desirable to use the same test with all children in a group, it is always better to match the type of test to the individual student's needs.

For example, a group may be tested using the paper edition of the *New Group Reading Test*. In this series there are four levels of test that are age-appropriate. The full range of scores will be available only if the student is given the correct test for his or her age. This will mean that for some students the test will be too difficult. In addition, some students with weak skills may feel very anxious about a group testing situation and be unable to engage with the tasks demanded by the test.

It would be better to use an individual assessment such as the *York Assessment of Reading for Comprehension* which allows the student to work with material which is matched to their reading ability so scores reflect both the age of the student and the difficulty of the material used for testing. Diagnostic information is enhanced; *YARC* gives scores for rate, error and comprehension. If a student's reading is very delayed the *YARC Early Reading* suite of tests may be used and in this case a reading age will be available and become the basis for tracking progress.

If a student will tolerate only a short test time, then a word level reading test may be used, such as the *Single Word Reading Test* (which is for students from age 6 to 16) or the *Diagnostic Test of Word Reading Processes* (for students from age 5 to 12).

Test administration

Advance preparation for the administration of any test is very important.

The teacher/administrator must read the manual and understand that the test environment must be strictly controlled and students prepared adequately for the test experience. The administration instructions must not be changed and students must not be helped to answer the test questions. However, it is permissible to help them with the method of responding and clarify any points about the process of the test session.

A key purpose of standardised tests is to find out what a student can do in order to help them achieve more. This needs to be conveyed to the students who must 'do their best' and work consistently through the test.

Close invigilation is necessary and the teacher/ administrator must periodically walk around the room checking that students understand how to complete the test.

Some tests are untimed and, if so, the teacher must make sure that enough time has been given to allow most students to complete the test; guidance is usually given about this in the test manual. Where a test is timed it is very important that the given timings are adhered to exactly.

Re-testing

A general guideline is that re-testing, even when equivalent forms are available, should be carried out at a six-month interval. The majority of schools will choose to test once a year with curriculum-linked tests (and to be most effective, at the same time of year) and possibly twice a year with a skills-based test. Ability tests such as *CAT* are usually administered every two or three years. Re-testing with the same test content should not be carried out within six months.

Where an intervention has been put in place and equivalent forms are available it is acceptable to test over a shorter timeframe; however, if the period is too short the test may not be able to measure the impact of the intervention. Also, the test must reflect the skills the intervention aims to develop. A minimum three-month period would seem appropriate in such situations.

Understanding test scores

The following information is common in many GL Assessment tests and features on the group and individual reports for tests such as the *Cognitive Abilities Test 4* and *New Group Reading Test* for both paper and digital versions.

The **SAS** is the most important piece of information derived from any standardised test. The SAS is based on the student's raw score which has been adjusted for age and placed on a scale that makes a comparison with a nationally representative sample of students of the same age across the UK. The average score is 100. The SAS is key to benchmarking and tracking progress and is the fairest way to compare the performance of different students within a year group or across year groups. See Appendix for more detailed information.

The **stanine** places the student's score on a scale of 1 (low) to 9 (high) and offers a broad overview of his or her performance.

Performance on a test can be influenced by a number of factors and the **confidence bands** are an indication of the range within which a student's scores lies. The narrower the band the more reliable the score, and 90% confidences bands are a high-level estimate.

The **Group Rank** (**GR**) shows how each student has performed in comparison to those in the defined group.

The **National Percentile Rank** (**NPR**) relates to the SAS and indicates the percentage of students obtaining any score or below. NPR of 50 is average. NPR of 5 means that the student's score is within the lowest 5% of the national sample and NPR of 95 means that the student's score is within the highest 5% of the national sample.

The **National Curriculum** (**NC**) **reading level** is based on teacher assessment collected when the test was developed. It is an estimate of the level the student has attained at the time the test was administered.

The **reading age** (or age-equivalent score) is the age at which a particular score is obtained by the average student based on the national sample.

New Group Reading Test (NGRT)

GL Assessment's New Group Reading Test is used very widely both for regular progress checking and in the pre- and post-intervention test for several national studies. This use has prompted a range of questions which this section attempts to address.

The NGRT is available in two editions:

- Paper tests – four age-appropriate levels for Years 1 to 11
- Digital, adaptive tests – two equivalent forms and each form is for students from age 7 to 16

The digital test adapts the test material to suit the performance of each student as they take the test. This means that students with high ability will be given material that is more challenging than that in the paper tests and vice-versa for students with very low ability. This makes the testing process more appropriate to the range of ability within any one group and should help students' engagement with the test.

NGRT and National Curriculum levels

The NC levels given in *NGRT* are based on teachers' assessment of their students' level of performance in reading at the time the test was administered. Based on a large standardisation sample (n = 11,640) an analysis of these data was carried out and levels assigned based on the raw score for the whole test.

This can only ever be an indication of a student's NC level as *NGRT* tests just two aspects of reading: sentence level reading through sentence completion and reading comprehension based on one, two or three short passages.

Teacher assessment of reading (or English) will be based on the student's performance across all activities. The benefit of the information from *NGRT* is that it can provide objective information based on a nationally standardised test; however, as stated this is for two aspects of reading only.

A difference in teacher assessment and the NC level offered by *NGRT* of one or two sub-levels does not signal an error in the test but means that in terms of the student's performance on the test they demonstrated skills in sentence completion and reading comprehension at a particular level. The teacher's assessment should be the main indicator of performance overall.

NGRT and the Reading Ability Scale (RAS)

To create the adaptive digital edition of *NGRT* all content was 'scaled', that is, put in order of difficulty so that individual questions in sentence completion could be administered in order of difficulty and passages (as a whole) could be administered in the same way.

The sentence completion part of *NGRT* adapts to the students' performance as they are answering questions; this then indicates which passage should be administered first. Passages are given as a whole and at the end of the first passage, the score indicates which passage should be given next; this may be a harder or an easier passage.

The same scale has been used to classify ages of students so that a student of, say 7 years and 6 months, is given, as the first sentence completion question, one that is at a difficulty level appropriate to an age that is slightly below 7 years 6 months. This allows the student to 'work up' or 'work down' to an appropriate level.

At the end of the test the RAS is recalculated based on sentence completion and passage comprehension and this is clearly shown on the Individual Report for Teachers (IRT)[1] and informs the reading age given to the student.

NGRT and Reading age-equivalent scores (reading ages)

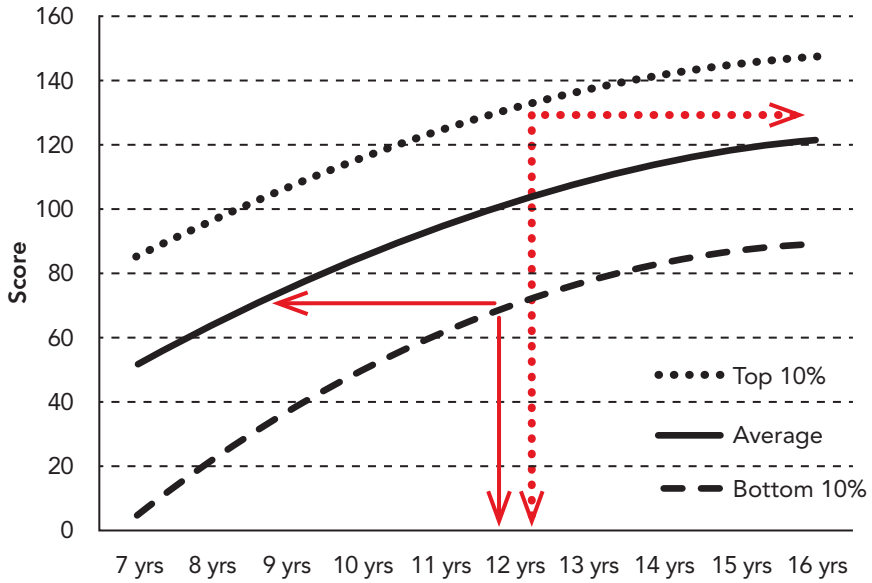
Reading ages are not the same as reading attainment. Age equivalents are derived from the **average raw scores** at **different age** points.

The graph below shows the relationship between raw scores and student age for a reading test. The solid black line shows the average scores, for example the average raw score for 7-year-olds is 50. Therefore any student with a raw score of 50 will have an age-equivalence or reading age of 7 years. The graph also shows the spread of raw scores within the age groups by displaying scores in the top 10th percentile (dotted line) and bottom 10th percentile (dashed line).

[1]Available at www.gl-assessment.co.uk in the section about NGRT Digital

For example, a 12-year-old student with a raw score of 70 and in the bottom 10th percentile will have an age-equivalent of 9 years (as the average raw score for 9 year olds is 70). The *solid red line* displays this relationship.

Figure 1 Score by age



For older students in secondary schools there is not much difference in the average raw score of 14-, 15- or 16-year-olds. Therefore a small increase in raw score can lead to big increases in age-equivalents.

In most cases it is not sensible to relate scores for students with above average ability to age- equivalents as age-equivalents by definition relate to an average. For example, the most we can say about a 12-year-old student with a raw score of 130 and in the top 10th percentile is that his age-equivalence is 16 years+, so this student is performing better than an average 16-year-old (*refer to the red dotted line*). Therefore there are issues with using age-equivalents and it is best to use standard age scores for measuring progress or monitoring trends.

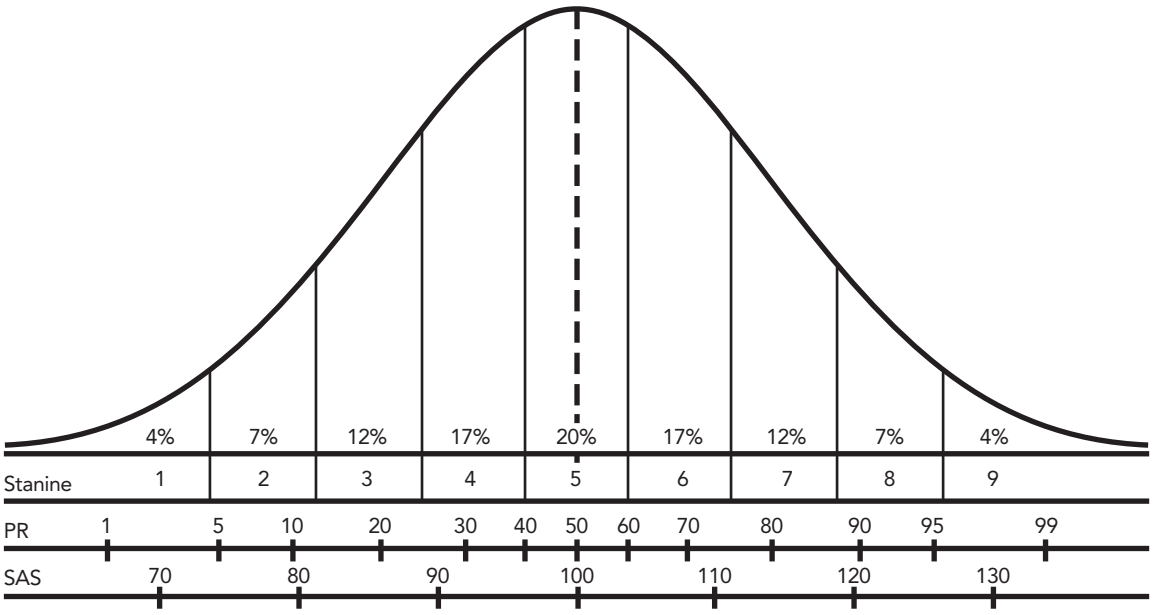
GL Assessment
May 2013

Appendix

One way to make a raw score more readily understandable would be to convert it to a percentage: for example, '33 out of 50' becomes 66 per cent. However, the percentage on its own does not tell us the average score of all the students or how 'spread out' the scores are, whereas standard age scores do relate to these statistics.

In order to provide a standard age (or standard score) scale, some tests are standardised so that the average standard age score for any age group is always 100: this makes it easy to tell whether a student is above or below the national average. The spread of scores (the 'standard deviation') is also set to plus or minus 15 points, so that for any age group about two-thirds of the students in the national sample will have a standardised score of between 85 and 115. Raw scores are converted to standard age scores that allow you to compare the level of cognitive development of an individual with the levels of other students in the same age group. The properties of standard age scores mean that approximately two-thirds of students in the age group score between 85 and 115, approximately 95 per cent score between 70 and 130, and over 99 per cent score between 60 and 140. Figure A shows the frequency distribution, known as the normal distribution, for standard age scores, stanines and percentiles.

Figure A The normal curve of distribution showing the relationships of stanines, national percentile ranks (PR) and standard age scores (SAS)



Standard age scores have three particular benefits, as described below.

- *They place a student's performance on a readily understandable scale*. As we have seen, standard age scores allow a student's performance to be readily interpreted. It is immediately deducible from the score itself that a verbal reasoning score of 95 indicates a level of performance just below the national average, but well within the average range.

- *An allowance can be made for the different ages of the students.* In a typical class the oldest students are very nearly 12 months older than the youngest. Almost invariably, older students achieve slightly higher raw scores in tests and examinations than younger students. However, standard age scores are derived in such a way that the ages of the students are taken into account by comparing a student *only* with others of the *same age*. An older student may in fact gain a higher raw score than a younger student, but have a lower standardised score. This is because the older student is being compared with other older students in the norm group. Students of different ages who gain the same standard age score have done equally well, with each being judged in relation to their standing among students of their own age.

- *Scores from different tests can be meaningfully added or compared.* Standardised scores for most tests cover the same range, from 60- to 140+. Hence a student's standing in, say, maths and English can be compared directly using standardised scores. It is not meaningful to add together raw scores from tests of different length or difficulty. However, should you wish to add *standardised* scores from more than one test – for example, in order to obtain a single overall measure of attainment – they can be meaningfully combined.

GL assessment
the measure of potential

www.gl-assessment.co.uk