Verbal Reasoning: Technical information

Issues in test construction

In constructing the *Verbal Reasoning* tests, a number of important technical features were carefully considered by the test constructors.

The questions themselves

These were all trialled before the final version of the tests were compiled, and some questions were rewritten or discarded because of their unsuitable level of difficulty, ambiguity, gender bias or other poor technical quality. Further details are given in the Technical information section below.

The consistency of the tests as a whole

There are two reasons why you can be reassured that the tests are reliable. First, because of the controlled test delivery and automatic computer marking, there is very little room for the test to be administered in different ways and thus to obtain different results. Second, the tests are very reliable in the sense that they would give very similar results to those which would theoretically be obtained by a very long verbal reasoning test, or repeated testing, with the pupils suffering no fatigue or lack of motivation. Further information on reliability is given below.

The validity of the tests

The test developers were careful to ensure that the tests actually measure verbal reasoning as described in the *Guidance and Information for* Teachers booklet. Also, if you were to compare scores from these tests with those from other verbal reasoning tests, you would find a fairly high level of agreement. Thus, the GL Assessment *Verbal Reasoning* tests can be trusted as a modern and reliable test for which the item types and the questions themselves have been carefully selected. Further information on validity is given below.

Practice and learning effects

It is likely that pupils who retake the *Verbal Reasoning* tests, particularly within a short time, will improve their scores by at least a few raw score points, on average. Research with other tests of a similar nature has found this to be the case. This improvement will be the result of increased familiarity with the general test situation and question formats, and also specific familiarity with the questions themselves. You should therefore be aware of this effect, and avoid treating any modest improvements in scores as evidence of a genuine improvement in reasoning ability.

No attempt has been made to provide specific figures for the average increment on retesting, since these could be very misleading. A low increase in a research study could mean that the tests are genuinely resistant to a practice effect, or it could simply show that pupils were less motivated when asked to retake the same test after only a short period of time, thus countering any increment due to practice. There could also be a wide variation in the extent to which individual pupils improve their scores. In any case, it is advisable for teachers who wish to re-administer a test to wait for a more substantial period, say a year or two, and then to administer another GL Assessment *Verbal Reasoning* test at a different level.

Development of the tests

The initial stage of specifying the content of the tests involved examining the range of question types that are included in the NFER's modern Verbal Reasoning Item Bank. Questions were then trialled in November 1991.

Test level	item types chosen for item trials	Number of new items written	Number of trial booklets compiled	Number of pupils each booklet was administered to	Number of questions chosen
VR8&9	14	103	2	320	65
VR10&11	16	121	2	250	75
VR12&13	17	133	2	320	85

The questions chosen not only an appropriate range of difficulty, but also the most statistically discriminating of the items trialled. Further, statistical tests of gender bias in individual questions were conducted, and those items were discarded in which either boys or girls did disproportionately better than the opposite sex when performance on the individual question was compared to performance on the test as a whole.

The time limits set for each test are considered to be the most appropriate balance between speed of working, attention span and test length, and take into account feedback from the schools participating in the trials. The various question types are arranged as far as possible in order of increasing difficulty, and the questions themselves are arranged in order of difficulty within each type.

Standardisation

The *Verbal Reasoning* tests were standardised in June 1992 using a national sample of schools that were randomly selected from the national register of maintained and independent schools. The sample was proportionately stratified by the following variables to give a more accurate representation of the country as a whole.

Test level	Region	LA type	School type	Number of schools	Size of year group	Number of pupils in each year group	Pupil age range	Average scores
VR8&9	Wales; North, Midlands and South of England	Metropolitan and non-metropolitan	Junior, Junior & Infants, Middle, Independent	174	3–30; 31–60; 61+ pupils	*Year 3:* 2,229 pupils (1,130 boys and 1,099 girls). *Year 4:* 2,951 pupils (1,551 boys and 14,00 girls).	7:09 to 9:09 statistically extrapolated to extend from 7:03 to 10:03.	*Year 3:* 34.6 out of 65 (53 per cent). *Year 4:* 44.7 out of 65 (69 per cent).
VR10&11				185		*Year 5:* 2,806 pupils (1,459 boys and 1,347 girls). *Year 6:* 2,644 pupils (1,317 boys and 1,327 girls).	9:09 to 11:09, statistically extrapolated to extend from 9:03 to 12:03.	*Year 5:* 37.6 out of 75 (50 per cent). *Year 6:* 46.9 out of 75 (63 per cent).
VR12&13			Grammar, Secondary Modern, Comprehensive and others, Independent	194	3–60; 61–120; 121+ pupils	*Year 7:* 2,471 pupils (1,252 boys and 1,219 girls). *Year 8:* 2,161 pupils (1,142 boys and 1,019 girls).	11:09 to 13:09, statistically extrapolated to extend from 11:03 to 14:03.	*Year 7:* 36.4 out of 85 (43 per cent). *Year 8:* 42.1 out of 85 (50 per cent)

Reliability

The reliability of a test is a measure of the extent to which a pupil's test scores would vary with repeated testing, assuming that there was no fatigue, learning or lack of motivation. The more consistent the scores, the higher the reliability.

K-R 20 reliability

VR8&9

The first estimate of reliability for *VR8&9* was derived from the Kuder-Richardson 20 formula (K-R 20) that measures the internal consistency of the test. This is 0.96 for both the Year 3 sample and the Year 4 sample. Based on the combined sample used to compute the standardised scores, the reliability is 0.97, which is felt to be suitably high for a modern verbal reasoning test. With the standardised scores having a standard deviation of 14.8, this gives a standard error of measurement (SEM) of 2.7.

The K-R 20 reliability was used to calculate the confidence bands. Adding and subtracting one SEM to or from 100, the average standardised score on *VR8&9*, gives a range from 97.3 to 102.7, and there is a 68 per cent chance (about 2 in 3) that the pupil's true score will be in this range.

If greater certainty is needed, then the size of the confidence band will be increased. For example, a 90 per cent confidence band, which corresponds to adding or subtracting 1.645 times the SEM to or from the pupil's score. In the above example of a score of 100, this would mean adding or subtracting 4.4 (giving a range of 95.6 to 104.4), and hence there is a 9 out of 10 (90 per cent) chance that the true score lies within the band.

Re-test reliability

A random sample of the schools that took part in the June 1992 standardisation was invited to administer the test again in June 1993. 306 pupils took the test on both occasions. The correlation between the raw scores was 0.81 and the correlation between the standardised scores was also 0.81. The mean raw score on the first administration was 38.3 and was 48.6 on the second, a rise of 10.3 ($p<0.001$). The mean standardised scores on these two occasions were 103.6 and 104.9, a rise of 1.3 ($p=0.007$).

VR10&11

The first estimate of reliability for *VR10&11* was derived from the Kuder-Richardson 20 formula (K-R 20) that measures the internal consistency of the test. This is 0.96 for both the Year 5 sample and the Year 6 sample. Based on the combined sample used to compute the standardised scores, the reliability is 0.96, which is felt to be suitably high for a modern verbal reasoning test. With the standardised scores having a standard deviation of 14.8, this gives a standard error of measurement (SEM) of 2.9.

The K-R 20 reliability was used to calculate the confidence bands. Adding and subtracting one SEM to or from 100, the average standardised score on *VR10&11*, gives a range from 97.1 to 102.9, and there is a 68 per cent chance (about 2 in 3) that the pupil's true score will be in this range.

If greater certainty is needed, then the size of the confidence band will be increased. For example, a 90 per cent confidence band, which corresponds to adding or subtracting 1.645 times the SEM to or from the pupil's score. In the above example of a score of 100, this would mean adding or subtracting 4.8 (giving a range of 95.2 to 104.8), and hence there is a 9 out of 10 (90 per cent) chance that the true score lies within the band.

Re-test reliability

A random sample of the schools that took part in the June 1992 standardisation was invited to administer the test again in June 1993. 317 pupils took the test on both occasions. The correlation between the raw scores was 0.88 and the correlation between the standardised scores was also 0.88. The mean raw score on the first administration was 36.9 and was 46.4 on the second, a rise of 9.5 ($p<0.001$). The mean standardised scores on these two occasions were 99.4 and 100.0, a rise of 0.6 ($p=0.101$).

VR12&13

The first estimate of reliability for *VR12&13* was derived from the Kuder-Richardson 20 formula (K-R 20) that measures the internal consistency of the test. This is 0.95 for the Year 7 sample and 0.96 for the Year 8 sample. Based on the combined sample used to compute the standardised scores, the reliability is 0.96, which is felt to be suitably high for a modern verbal reasoning test. With the standardised scores having a standard deviation of 15.0, this gives a standard error of measurement (SEM) of 3.0.

The K-R 20 reliability was used to calculate the confidence bands. Adding and subtracting one SEM to or from 100, the average standardised score on *VR12&13*, gives a range from 97.0 to 103.0, and there is a 68 per cent chance (about 2 in 3) that the pupil's true score will be in this range.

If greater certainty is needed, then the size of the confidence band will be increased. For example, a 90 per cent confidence band, which corresponds to adding or subtracting 1.645 times the SEM to or from the pupil's score. In the above example of a score of 100, this would mean adding or subtracting 4.9 (giving a range of 95.1 to 104.9), and hence there is a 9 out of 10 (90 per cent) chance that the true score lies within the band.

Re-test reliability
A random sample of the schools that took part in the June 1992 standardisation was invited to administer the test again in June 1993. 290 pupils took the test on both occasions. The correlation between the raw scores was 0.88 and the correlation between the standardised scores was 0.87. The mean raw score on the first administration was 36.2 and was 45.7 on the second, a rise of 9.5 ($p<0.001$). The mean standardised scores on these two occasions were 9.4 and 12.2, a rise of 2.8 ($p<0.001$).

Validity

Test validity may take several forms, based either upon the content of the test or upon the degree of association with other similar tests, the latter usually taking the form of 'concurrent' or 'predictive' validity.

Content validity
Verbal reasoning tests principally assess inferential and deductive skills through a variety of question types that are mainly concerned with the production of, use of, and relationships between words; also, the manipulation of letters or numbers may be included. An examination of the item types in the GL Assessment *Verbal Reasoning* tests that these processes are reflected in the individual questions of the test.

Concurrent validity
This was assessed by comparing scores from the *Verbal Reasoning 8&9*, *10&11* and *12&13* with scores from the older *Verbal Reasoning Test BC*, *D* and *EF* respectively (the equivalent tests in the original series of *nferNelson verbal reasoning tests). A sample of 327 (*VR8&9*), 315 (*VR10&*11) and 289 (*VR12&13)* pupils took the two tests, with about half taking the new test first and the other half taking the original test first. These two tests were taken within no more than a few days of each other.

For *VR8&9* the correlation between the raw scores was 0.82 and between the standardised scores was 0.78. For *VR10&11* the correlation between the raw scores was 0.86 and between the standardised scores was 0.87. For *VR12&13* the correlation between the raw scores was 0.89 and between the standardised scores was 0.86. These correlations are satisfactorily high, bearing in mind the following two factors:

- a typical re-test correlation for a verbal reasoning test, when re-administered in the space of a few days, is around 0.90;
- some of the item types are different in the original and the newer tests.

Predictive validity
Although time did not allow any research to this effect to be conducted during the development of these *Verbal Reasoning* tests, studies over the years with a number of other verbal reasoning tests taken by older pupils have found these tests to be relatively good predictors of subsequent academic performance. For example, verbal reasoning tests administered to eleven-year-old pupils have been found to correlate well with examination performance in the second and third years of secondary education and, more recently, with pupils' overall GCSE performance. It is hoped that work on predictive validity will be carried out over the next few years. The publishers would be pleased to hear from any teachers who are eventually able to provide suitable evidence, that is, who test a group of pupils and then, at a later date, obtain measures of attainment such as National Tests or standardised scores on reading or mathematics tests. If this test has succeeded in measuring key reasoning processes, then scores on it should predict future success in learning.