

Non-Verbal Reasoning: Technical information

Issues in test construction

The questions themselves

The questions in the tests result from trialling a much larger number of items. Questions were discarded that were at an unsuitable level of difficulty and that proved to be biased against either boys or girls. Questions were also eliminated that did not discriminate well between high and low scorers, possibly because they needed too high a level of spatial ability, visual discrimination or general mathematical knowledge.

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 give a very similar result to those which would theoretically be obtained by a very long non-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 non-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 non-verbal reasoning tests, you would find a fairly high level of agreement. So the GL Assessment *Non-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.

Non-verbal reasoning and 'IQ'

Although the *Non-Verbal Reasoning* tests certainly assess mental processes that are central to many people's notions of 'intelligence', it is strongly recommended that you avoid using terms like 'IQ' or 'intelligence' to refer to the test scores. It is more accurate to use the term 'IQ' only for the results of individually administered tests that assess a range of verbal and non-verbal abilities. Also, terms like 'IQ' and 'intelligence', loosely employed, may suggest somewhat misleadingly that the test is measuring something that is fixed and completely unrelated to the pupil's home environment.

Practice and learning effects

It is likely that pupils who retake the *Non-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. The 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 test is 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 Non-Verbal Reasoning* test at a different level.

Development of the tests

- *NVR8&9*
These questions were developed during the late 1980s, on the basis of an extensive survey of the relevant research literature. All the questions were part of a larger set that was trialled several times on large, representative samples of pupils of the appropriate ages. The 42 questions chosen reflect not only an appropriate range of difficulty, but also the most statistically discriminating of the items trialled.
- *NVR10&11*
These questions originate from three sources. Most were developed during the late 1980s, on the basis of an extensive survey of the relevant research literature. All the questions were part of a larger set that was trialled several times on large, representative samples of pupils. In this way, ambiguous or biased questions were eliminated and accurate measures of question difficulty were obtained for the rest of the set. A small group of questions were adapted from the *NFER Non-Verbal Test BD*, which this test was intended to supersede. Finally, some new items were created. All these questions were then retrialled with representative samples of the appropriate two year groups, in order to produce up-to-date measures of their difficulty. The 54 questions chosen reflect not only an appropriate range of difficulty, but also the most statistically discriminating of the items.

- *NVR12–14*

These questions originate from three sources. Most were developed during the late 1980s, on the basis of an extensive survey of the relevant research literature. All the questions were part of a larger set that was trialled several times on large, representative samples of pupils. In this way, ambiguous or biased questions were eliminated, and accurate measures of question difficulty were obtained for the rest of the set. A small group of questions were adapted from the NFER *Non-Verbal Test DH*, which this test was intended to supersede. Finally, some new items were created. All these questions were then retrialled with representative samples of the appropriate two year groups, in order to produce up-to-date measures of their difficulty. The 52 questions chosen reflect not only an appropriate range of difficulty, but also the most statistically discriminating of the items.

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 questions are arranged in order of difficulty within each type, and the instructions, examples and practice questions are those that proved satisfactory in the trials.

Standardisation

The *Non-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
NVR8&9	Wales; North, Midlands and South of England	Metropolitan and non-metropolitan	Junior, Junior & Infants, Middle, Independent	204	3–30; 31–60; 61+ pupils	Year 3: 3,214 pupils (1,617 boys and 1,597 girls).	7:09 to 9:09 statistically extrapolated to extend from 7:03 to 10:03.	Year 3: 25.0 out of 42 (60 per cent).
				Year 4: 3,061 pupils (1,574 boys and 1,487 girls).		Year 4: 29.4 out of 42 (70 per cent).		
NVR10&11			192	Year 5: 2,704 pupils (1,391 boys and 1,313 girls).	9:09 to 11:09, statistically extrapolated to extend from 9:03 to 12:03.	Year 5: 30.2 out of 54 (56 per cent).		
			Year 6: 3,069 pupils (1,530 boys and 1,539 girls).	Year 6: 35.2 out of 54 (65 per cent).				
NVR12–&14			Grammar, Secondary Modern (High), Comprehensive and others, Independent	317	3–60; 61–120; 121+ pupils	Year 7: 2,905 pupils (1,390 boys and 1,515 girls).	11:09 to 14:09, statistically extrapolated to extend from 11:03 to 15:03.	Year 7: 28.3 out of 52 (54 per cent).
			Year 8: 2,768 pupils (1,354 boys and 1,414 girls).	Year 8: 31.2 out of 52 (60 per cent).				
			Year 9: 1,916 pupils (980 boys and 936 girls).	Year 8: 33.4 out of 52 (64 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

NVR8&9

The first estimate of reliability was derived from the Kuder-Richardson 20 formula (K-R 20) that measures the internal consistency of the test. This is 0.932 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 0.935, which is felt to be suitably high for a modern non-verbal reasoning test. With the standardised scores having a standard deviation of 14.8, this gives a standard error of measurement (SEM) of 3.8.

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 NVR8&9, gives a range from 96.2 to 103.8, 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 6.3 (giving a range of 93.7 to 106.3), 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. 292 pupils took the test on both occasions. The following statistics relate to re-testing this sub-sample:

- The correlation between the raw scores was 0.67 and the correlation between the standardised scores was 0.69.
- The mean raw score on the first administration was 26.8 and was 33.6 on the second, a rise of 6.7 ($p < 0.001$).
- The mean standardised scores on these two occasions were 102.8 and 107.3, a rise of 4.5 ($p = 0.001$).

NVR10&11

The first estimate of reliability was derived from the Kuder-Richardson 20 formula (K-R 20) that measures the internal consistency of the test. This is 0.93 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 also 0.93, which is felt to be suitably high for a modern non-verbal reasoning test. With the standardised scores having a standard deviation of 14.8, this gives a standard error of measurement (SEM) of 3.9.

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 NVR10&11, gives a range from 96.1 to 103.9, 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 6.4 (giving a range of 93.6 to 106.4), 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. 297 pupils took the test on both occasions. The following statistics relate to re-testing this sub-sample:

- The correlation between the raw scores was 0.79 and the correlation between the standardised scores was 0.80.
- The mean raw score on the first administration was 30.8 and was 38.6 on the second, a rise of 7.8 ($p < 0.001$).
- The mean standardised scores on these two occasions were 105.8 and 109.9, a rise of 4.1 ($p = 0.001$).

NVR12-14

The first estimate of reliability was derived from the Kuder-Richardson 20 formula (K-R 20) that measures the internal consistency of the test. This is 0.91 for the Year 7 sample, 0.91 for the Year 8 sample and 0.92 for the Year 9 sample. Based on the combined sample used to compute the standardised scores, the reliability is 0.92, which is felt to be suitably high for a modern non-verbal reasoning test. With the standardised scores having a standard deviation of 14.8, this gives a standard error of measurement (SEM) of 4.2.

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 NVR12-14, gives a range from 95.8 to 104.2, 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 6.9 (giving a range of 93.1 to 106.9), 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. 284 pupils took the test on both occasions. The following statistics relate to re-testing this sub-sample:

- The correlation between the raw scores was 0.81 and the correlation between the standardised scores was 0.80.
- The mean raw score on the first administration was 29.7 and was 35.3 on the second, a rise of 5.6 ($p < 0.001$).
- The mean standardised scores on these two occasions were 101.9 and 106.6, a rise of 4.7 ($p = 0.001$).

Validity

The validity of a test is the extent to which it measures what it is intended to measure. Evidence of a test's validity can be gathered in various ways.

Content validity

For tests like the *Non-Verbal Reasoning* tests, the most important evidence of validity is the way in which the content of the tests reflect research evidence as to the nature of the relevant ability, and the extent to which the need for irrelevant abilities has been successfully eliminated from the questions.

Concurrent and predictive validity

Another way in which validity evidence is gathered is by assessing the degree of relationship between test scores and other measures of the pupils' abilities and attainment. This can involve either other measures that are available at the time of testing, known as concurrent evidence, or measures that are taken some time afterwards, known as predictive evidence.

This was assessed by comparing scores from *Non-Verbal Reasoning 8&9, 10&11* and *12–14* with scores from the older *Verbal Reasoning Test BD, DH* and *DH* respectively (the equivalent tests in the original series of *Nfer-Nelson* non-verbal reasoning tests). A sample of 331 (*NVR8&9*), 318 (*NVR10&11*) and 317 (*NVR12–14*) 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 *NVR8&9* the correlation between the raw scores was 0.76 and between the standardised scores was 0.74. For *NVR10&11* the correlation between the raw scores was 0.81 and between the standardised scores was 0.81. For *NVR12–14* the correlation between the raw scores was 0.84 and between the standardised scores was 0.80. These correlations are satisfactorily high, bearing in mind the following two factors:

- a typical re-test correlation for a non-verbal test, when re-administered in the space of a few days, is around 0.88;
- some of the item types are different in the original and the newer tests.