Technical Information Ireland

Trialing of CAT4

The initial development work involved in developing the CAT4 was conducted by GL Assessment in the UK in 2009 and 2010. Small scale trials were conducted in autumn 2009 to check some of the new questions being developed for the *CAT4* Spatial Ability Battery. Three versions of the new spatial test were created and were trialed with approximately 850 students in Years 4, 6, 8 and 9. Results from this study were used to develop further spatial questions for the main trials.

The main trials of all the questions in all four batteries of *CAT4* were carried out in autumn 2010. The numbers of students taking part in the trials from Years 6 to 10 were: 2028, 1870, 2179, 2114 and 8191 respectively.

For the trials, 24 test booklets were created; that is six test booklets for each CAT4 Level. All students took Verbal Classification and Figure Recognition plus two of the remaining six test types, so that all items were taken by at least 300 students. Some of the questions were duplicated in booklets across CAT4 Levels.

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 sexes, once each sex's general level of performance was taken into account. This information was then used to select and order the sequences of questions for the final standardisation version of *CAT4*.

Standardisation of CAT4 in Ireland (2012)

Irish age-based norms for the CAT4 were derived from the administration of four levels of the test in the Spring of 2012 to students in 5th and 6th class in random samples of primary schools (levels D and E) and to students in First Year to Fifth Year in random samples of second-level schools (Levels E, F, and G). This work was conducted by EdEv Ltd (A Dublin City University campus company) and GL Assessment in the UK (the publisher of CAT4).

Sampling Frame Levels D and E (5th and 6th Class)

- 3165 primary schools were identified from the Department of Education and Skills (DES) list of schools 2010/11.
- 919 junior schools were excluded as they had no 5th and 6th classes. A further 55 schools were excluded as they were involved in previous standardisations involving CAT3.
- The remaining 2191 schools were explicitly stratified by size (large, small). This school list was implicitly stratified by DEIS status, gender mix and school size.
- The school sample was selected with Probability Proportional to Size (PPS). The Measure of Size (MOS) was number of pupils in 6th class.
- These schools were contacted in the autumn of 2011 and invited to participate in the standardisation. Schools that agreed to participate forwarded a list of classes within each year. From this list one 5th class and one 6th class from each school were selected at random.

Achieved Samples for Levels D and E (5th and 6th Class)

- The effective population size for 5th was 53,450.
- The effective population size for 6th was 50,798.
- The total number of schools selected was 60 and the total number of pupils in the selected classes was 1260 (5th, Level D) and 1045 (6th, Level D) and 1182 (6th, Level E).
- The achieved sample for schools was 49 or 82%. The numbers of pupils that participated in the standardization was:

  1007 in 5th class at Level D,
  750 in 6th class at Level D,
  899 in 6th class at Level E.

Sampling Frame Levels E, F and G (First to Fifth Year)

- 728 post-primary schools were identified from the DES list of schools 2010/11.
- 136 schools were excluded from the sampling frame. Of these 92 were excluded because they had participated in a previous standardisation of the CAT; the other 44 schools were excluded because the school had no students in one or more of 1st, 2nd, 3rd and 5th Years.
- Thus the Sampling Frame included the remaining 592 schools. This list was implicitly stratified by school type (Secondary, Vocational, Community/Comprehensive), DEIS status and school size.

The sample of schools (60) was selected with PPS. The MOS was number of students in First Year.

- These schools were contacted in the autumn of 2011 and invited to participate in the standardisation. Schools that agreed to participate forwarded a list of classes within each year. From this list one class from each year level was randomly selected.

Achieved Samples for Levels E, F and G (First to Fifth Year)

- The effective population size for each of the years was:

  49,540 First Year
  48,829 Second Year
  46,902 Third Year
  24,740 Fourth Year (TY)
  55,119 Fifth Year

- The total number of schools selected was 60 and the total number of students in First to Fifth Year in the selected classes was 1522, 1505, 1490, 1239, 1493 respectively.
- The achieved sample of schools was 45 or 75%. The numbers of pupils that participated in the standardization was:

  971 First Year Level E
  952 Second Year Level F
  839 Third Year Level F
  718 Fourth Year Level G
  809 Fifth Year Level G.

CAT4 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 was 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 Cronbach's Alpha formula which produces values ranging from 0 to 1. Values above 0.80 are considered to be very satisfactory. The reliability values for the various *CAT4* batteries based on the Irish standardization data are given in the table below, and all show that the tests are very reliable.

	CAT4 Reliability				
CAT4 *Level*	*Verbal Reasoning Battery*	*Quantitative Reasoning Battery*	*Non-verbal Reasoning Battery*	*Spatial Ability Battery*	*Overall* CAT4
D	0.89	0.90	0.88	0.87	0.96
E	0.89	0.88	0.86	0.87	0.95
F	0.90	0.87	0.84	0.88	0.95
G	0.91	0.86	0.83	0.88	0.95
Average D–G	0.90	0.88	0.85	0.87	0.95

Standard Error of Measurement

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. For example, The SEM for the Level D Verbal Reasoning Battery is 5.1, which indicates that there is a 68 per cent chance that the student's true verbal Standard Age Score (SAS) will be in the range +/- 5.1. For example, for an average-performing student with a verbal SAS of 100, there is a 68 per cent chance that his or her true Verbal Reasoning score lies in a range from 94.9 to 105.1.

	CAT4 SEM				
CAT4 *Level*	*Verbal Reasoning Battery*	*Quantitative Reasoning Battery*	*Non-verbal Reasoning Battery*	*Spatial Ability Battery*	*Overall* CAT4
D	5.1	4.8	5.2	5.4	3.1
E	5.0	5.2	5.7	5.4	3.2
F	4.8	5.4	6.1	5.2	3.3
G	4.4	5.7	6.2	5.3	3.3
Average D–G	4.8	5.3	5.8	5.3	3.2

However, most tests show the 90% chance or confidence bands. For values around the average, the 90% confidence band is as follows:

	CAT4 90% Confidence Band				
CAT4 *Level*	*Verbal Reasoning Battery*	*Quantitative Reasoning Battery*	*Non-verbal Reasoning Battery*	*Spatial Ability Battery*	*Overall* CAT4
Average D–G	+/- 8	+/- 9	+/- 10	+/- 9	+/- 5

For example, for an average-performing student with a Verbal Reasoning SAS of 100, there is a 90 per cent chance that the true score for Verbal Reasoning lies in a range from 92 to 108.

Evaluating Differences Between CAT4 Scores

The evaluation of a difference between two scores, whether scores on two different tests or scores on the same test on two occasions, has to be a three-stage process.

1. Statistical significance of differences

First, it needs to be decided if the difference is large enough to be considered as 'real' rather than being just a result of having imprecisely measured the two scores. This depends upon the test reliability of each of the two scores and hence, the 'noise' around each one.

The measurement error when calculating a difference between two scores is evaluated using a coefficient called the standard error of measurement difference (SEM_{diff}).

The SEM_{diff} for *CAT4* scores is approximately 7 standard score points. Consequently, if two scores are more than 7 SAS points apart, it is 68% likely that they are real and if they are more than 11 points apart, the likelihood is 90% that the difference is a real one.

2. Rarity of differences

Second, if the difference is 'real' or statistically significant, then the unusualness or rarity of the difference has to be evaluated. A significant difference can sometimes be very common. For example, if you use a millimetre ruler to measure a child's height when s/he is seven and then again when s/he is eight, the difference between these two heights can be measured very accurately to within two millimetres. Therefore 'real' or statistically significant differences will be very common in a sample of children because the difference between the heights is likely to be substantially greater than two millimetres in almost all cases.

The spread of difference in scores can be determined either directly from the data or by a formula that takes into account the spread of scores on each test and the correlation between the two sets of scores. If the sample size is large enough, the two methods will produce very similar results; this was the case for the standardisation of *CAT*. The formula used is:

$$SE_{diff} = \sqrt{(SD_1^2 + SD_2^2 - 2r_{12} SD_1 SD_2)}$$

Where SD_1 and SD_2 are the standard deviations of the scores on each test and r_{12} is the correlation between the two tests.

When looking at differences between a student's scores on the same battery on two occasions (e.g. Verbal in First Year and Verbal in Second Year), the table below can be used [1]. For example, a score increase of 11 SAS points or more will occur with between 10 and 15 per cent of children, but a decrease of 17 or more points will occur with only the most extreme 5 per cent.

Difference in SAS Scores from first to second occasion	Percentage of students obtaining this extent and direction of difference
Increases by >16	5%
Increases by >12	10%
Increases by >9	15%
Decreases by >9	15%
Decreases by >12	10%
Decreases by >16	5%

[1] The figures in the table have assumed a mean correlation of 0.8 between the two occasions.

When looking at score differences between different batteries (e.g. Quantitative and Non-verbal), this table should be used instead[2]. The SAS score differences are larger in this situation because the two measures are of different underlying mental processes, so tend to be less highly correlated than two scores on the same test.

Difference in SAS Scores from Battery 1 to Battery 2	Percentage of students obtaining this extent and direction of difference
Higher by >19	5%
Higher by >15	10%
Higher by >12	15%
Lower by >12	15%
Lower by >15	10%
Lower by >19	5%

[2] The figures in the table have assumed a mean correlation of 0.7 between pairs of batteries.

3. Practical significance of differences

Finally, it needs to be remembered that a difference between two batteries which occurs commonly in the general population is not necessarily insignificant. It can indicate a real, albeit common, difference between the development of the cognitive abilities underlying the two battery scores, with implications for the ways in which the student concerned is likely to progress academically. Such differences need to be interpreted in the light of all that is known of a student's background and educational record. For example, students who have a background of poor socio-economic and educational opportunities who gain higher scores for non-verbal reasoning than for verbal reasoning may not have any real difference between their abilities to reason with words and with shapes. Instead, they may not have had the chance to acquire the basic reading and word knowledge needed to perform well on the verbal tasks. On the other hand, if they have good socio-economic and educational backgrounds, then the score difference may suggest that there is a genuine difference in abilities to think with words and with shapes.

Gender Differences on CAT4

The table below provides the mean SAS scores and standard deviation for each of the *CAT4* batteries by gender.

	CAT4 *level*		*Verbal Reasoning Battery*	*Quantitative Reasoning Battery*	*Non-verbal Reasoning Battery*	*Spatial Ability Battery*	*Overall* CAT4
Female	D	Mean	99.72	98.58	99.46	99.23	99.23
		Std. Deviation	14.685	14.558	15.006	15.082	12.67
	E	Mean	101.09	99.65	101.17	100.25	100.49
		Std. Deviation	14.558	13.723	14.382	14.100	11.88
	F	Mean	100.19	99.02	101.02	99.31	99.86
		Std. Deviation	15.496	14.295	15.031	15.084	12.81
	G	Mean	100.50	98.42	100.80	99.29	99.72
		Std. Deviation	15.037	13.389	14.921	14.741	12.06
	Total	Mean	100.39	98.94	100.63	99.54	99.84
		Std. Deviation	14.942	13.999	14.835	14.740	12.35
Male	D	Mean	100.43	102.27	100.89	101.53	101.23
		Std. Deviation	15.028	14.996	15.029	14.628	12.51
	E	Mean	98.61	101.08	98.77	100.29	99.57
		Std. Deviation	15.592	16.185	15.492	15.779	13.56
	F	Mean	99.96	101.45	99.13	100.94	100.30
		Std. Deviation	14.623	15.330	14.571	14.461	12.35
	G	Mean	99.14	103.28	99.51	102.11	100.80
		Std. Deviation	15.206	16.364	15.582	15.864	13.33
	Total	Mean	99.57	101.89	99.59	101.12	100.44
		Std. Deviation	15.129	15.678	15.155	15.143	12.93

It is clear from the table that across the four levels of the CAT4, female and male SAS scores are very similar. That said, it is noticeable that, with the exception of Level D, females score slightly higher than males on the Verbal Reasoning Battery while at all levels of the test, males score higher in Quantitative Reasoning. The largest difference occurs for Quantitative Reasoning at Level G, where males performed on average almost five SAS points higher than their female counterparts. However, even in this case, when the standard error is taken into account, the two means are not statistically different.

Relationship Between CAT3 and CAT4 Scores

A study was carried out in the UK comparing the national distribution of *CAT3* and *CAT4* standard scores in autumn 2011 for *CAT* Level D. Results show that there is no significant difference. So, for example, a student getting a verbal SAS of 90 on *CAT3* is also likely to obtain a verbal SAS of 90 using *CAT4*. This is not surprising as the national averages of SAS scores based on a database of over 250,000 students who use Level D every year have not changed significantly in each of the last ten years. The national average *CAT3* standard score was 100 back in 2001 and the average standard score for both *CAT3* and *CAT4* Level D tests in 2011 was approximately 100.

CAT4 Paper–Digital Comparison Study

Two studies were conducted in the UK to see if there was a difference in the way students scored between the paper and digital editions of *CAT4*.

- The overall numbers of students doing the digital and paper versions in the standardisation sample were large. This allowed a study to be undertaken looking at the relative difference in scores between those students doing paper and digital editions during the *CAT4* standardisation.
- The second study, also in autumn 2011, looked at the results of an equivalence study conducted in three year groups. Around 1,300 students in this study did both the paper and digital versions of the *CAT4* Non-Verbal Battery for Levels A, B and E. To reduce practice effects, around half the students completed the paper edition first followed by digital while the other half took the digital edition first followed by the paper edition.

The results of both studies have shown small differences in scores, with students completing the paper edition scoring slightly higher on average than on the digital edition. For example, the Non-verbal Reasoning Battery Level E paper raw score is, on average, half a point higher than for the digital edition and around 1 point higher for Level B.

The normative scores have therefore been adjusted to take into account any differences in the way students respond digitally or on paper.

Linking Performance on CAT4 and Performance on the Irish State Examinations

A study linking performance on CAT4 and performance on the State Examinations (Junior and Leaving Certificates) is being planned for 2016. The outcomes from this *Indicators Project* will be included here when the study is completed.