OCR Level 7 Diploma
Ian Reed

Principles of Psychometric and Educational Assessment.

"Assessments are needed to diagnose, to plan intervention, to inform school/college policies, to support claims for funding, to justify special arrangements in examinations, and more."

(ed. Backhouse & Morris, 2005)

Introduction.

Psychometric testing has it's origins in attempts to measure intelligence, with the first tests devised by Francis Galton (1822 – 1911) - see appendix 4. As a look into intelligence tests will give us a general feel for psychometric and educational assessment, this will be discussed first.

Concepts of Intelligence.

When one thinks of 'intelligence' and what it comprises, it is fairly certain that no two people will arrive at the same conclusions. Indeed, it is very easy to come up with a myriad number of concepts with different levels of interrelatedness and correlation.

A look at how human intelligence/human mental abilities are actually measured will help with our understanding of the term 'intelligence'.

g	→	Verbal Comprehension →	
		Vocabulary (general intelligence)	
		Similarities	
		Information Comprehension	
		Perceptual Comprehension	Picture completion Block design Picture arrangement Matrix

reasoning		
Processing Speed	→	Symbol search Digit-symbol coding
Working Memory	→	Letter-number sequencing Digit span Arithmetic

Hierarchy of mental ability - The Wechsler Adult Intelligence Scale III (Deary, 2001)[1]

The most fascinating thing about intelligence tests is also the most surprising; looking at the right column, it's easy to pick out some tasks that you believe that you would better at and some where you would not score as well. We also pick out special mental abilities that a person may have in contrast with their other abilities – "they're good at figures, but they have no common sense!" Indeed, looking at these tasks one finds it hard to find relations between them, or that being good at one task might mean being worse at another. In fact, none of these predictions is correct. Every one of the thirteen tasks has a positive correlation[2] with every other one, that is to say that a good score at one test tends to bring with it good scores from the others. In 1993 an American psychologist John Carroll published his work studying over four hundred datasets (basically every human mental testing data from the last century) and found positive correlation between the eight broad types of intelligence that make up g – see appendix 5.

On the semi-popular fringes of scientific psychology (Dearry, ibid), Howard Gardner has written about 'multiple intelligences' and has suggested, against the evidence, that there are many forms of mental ability and that they are unrelated. This view, not popular within scientific psychology has unsurprisingly become popular in education.

The relevance to our discussion concerning assessing pupils is to the concepts general and specific learning difficulties. Individuals with general learning difficulties show the same positive correlation between tasks – they perform consistently low. An individual with specific learning difficulties however, will not share the correlation, their results can be radically different between tests – producing what is known as a 'spiky' profile.

[1] The Wechsler is a closed test, it is only available to psychologists. The WRIT intelligence test overview, which is available to schools, can be seen in appendix 6.
[2] This will be discussed later as 'Internal consistency reliability' – see page 8.

Keep the diagram in appendix 5 in your mind (and the positive correlations associated with it) as we move on to other forms of psychometric and educational assessment – imagine, for example, reading instead of g in the centre and the mental abilities outer ring being replaced by the contributing abilities/skills for reading.

Aims of Assessment.

Before we undertake an investigation into the assortment of assessment materials, it would be worthwhile to look at the overall aims of assessment, as identified by Reid (1996);

- Identification of the learner's general strengths and weaknesses.
- An indication of the learner's current level of performance in attainments.
- An explanation for the learner's lack of progress.
- Identification of the aspects of the learner's performance in reading, writing and spelling, which may typify a 'pattern of errors '.
- Identification of specific areas of competence.
- An understanding of the student's learning style.
- An indication of aspects of the curriculum which may interest and motivate the learner.

A child's strengths and weaknesses may be obvious to the class teacher, who has day-to-day contact with the pupil, but as we will discuss in the next section; an identified weakness in reading does not lead on to an identification of the *reasons behind* that weakness. Assessments can unearth a pattern of difficulties that can lead to a structured program of teaching to help the pupil.

Difficulties Made Visible.

The specialist teacher, with their goal of preventing difficulties in learning, requires the elimination of equivocation when dealing with terms such as 'reading difficulty'. In our use of ordinary language, someone may be described as being "not very good at reading". This term can be used in numerous situations to describe quite different things, for example, it can be used when referring to a child that has difficulties with phonological awareness or when a pupil with poor sequencing is spoken of.

It will be useful here to borrow two terms from philosophy to aid us in our understanding, ultimate and proximate. The terms ultimate and proximate are used for describing causation. The ultimate cause can be seen as the higher order reason whereas the proximate cause "is an

event which is *closest* to, or immediately responsible for causing, some observed result" (Guttenplan, 2002). These terms will be applied when discussing difficulties. An ultimate difficulty is one in which we employ in our use of ordinary language, for example 'spelling' (a spelling difficulty) or 'writing' (a writing difficulty). A proximate difficulty is a more in-depth description that is related with the ultimate difficulty. The application of these terms will become clearer with an example;

A problem in reading can be seen as an ultimate difficulty, whereas a problem with phonological awareness is a proximate difficulty. The proximate difficulty causes the ultimate difficulty[3] Different proximate difficulties and combinations of them may give rise to the same ultimate difficulty.

When seen from this standpoint, the idea that someone has a problem with 'reading' or 'spelling' does not tell us anything that will lead to any successful intervention, a difficulty playing the piano will not be alleviated by simply more playing. The following table gives the relationship between ultimate and proximate difficulties for three different children.

2. It is worth considering the rules of implication; if $p \supset q$ is true (if p then q), then $q \supset p$ (if q then p) is not necessarily true, phonological awareness may cause/imply a difficulty in reading, but it is not the case that a difficulty in reading implies a difficulty with phonological awareness – the current thinking that synthetic phonics will improve reading ignores this. Logical implication is discussed in more detail in appendix 1.

Ultimate difficulty	SpLD	Proximate difficulties
Child A		
Reading	Dyslexia	Comprehension Automaticity Phonological awareness Self esteem
Child B		
Reading	Dyslexia	Sequencing Short-term memory Phonological awareness Malapropisms Disorganisation
Child C		
Clumsy	Dyspraxia	Impulsivity Automaticity Disorganisation Handwriting

(Inclusion development Program, 2007, National Strategies)

This table shows how a diagnosis of a specific learning difficulty fits in-between the ultimate and the proximate difficulties, a misunderstanding of the proximate difficulties can often lead to class teachers to the wrong conclusions "he is really disorganised, I think he's dyslexic". Here we can see that the proximate difficulty of disorganisation fits into the profile of the dyslexic and the dyspraxic child (a clear misuse of the rules of implication – see footnote 2).

A problem with kicking a ball can be seen as an analogy here, two children may share the ultimate difficulty (kicking a ball) but have different proximate difficulties that cause it, for example poor muscular coordination or poor spatial awareness. The ultimate difficulty is open for all to see whereas the underlying proximate difficulty is hidden from view.

It is through both psychometric and educational assessments that the proximate difficulties can be made visible so that a profile can be accumulated and a diagnosis of a specific learning difficulty (SpLD) can be made, resulting in a carefully planned and appropriate teaching programme.

Assessments.

Assessments can be carried out in different ways, an ipsative assessment is one in which the outcome is measured against an earlier score acquired by the same child. A criterion-referenced test will measure a pupil against a given criterion (example). A norm-referenced test will "compare learners performance with that of other individuals in the normative sample" (Boyle & Fisher, 2008).

Norm Referenced – this gives the test score a position in a predefined population N (see below).

Criterion Referenced – this gives a test score based upon the expected results with regard to the subject matter, an example would be a spelling test given by a teacher where he/she has already set an expectation with regards what constitutes a good score.

Ipsative – this gives a test score relative to the pupils own ability, pupil A scored X last month and now they have scored Y.

It is important to note here that, "planning for a diagnostic assessment should take account of a wide range of background information about the individual's developmental and educational history" (Waine & Kime, 2005, p.30). It is necessary to gain information about the pupil to get a better idea of the (ultimate) difficulties he/she may have. This information can be collected via questionnaires for the class teacher, the parents and, when appropriate, the pupil. Other sources of information can be samples of the child's work and any results of previous assessments or screenings (for example Dyslexia Screener, Smith & Turner, 2004)[4]. This background information is essential as there is a wide collection of assessment material available and the specialist teacher will have to choose an appropriate range of assessments to use. Assessments can be broadly separated into two groups, standardised and non-standardised. Non-standardised tests will be looked at first;

Non-Standardised Assessments.

Non-standardised assessments can viewed as containing two types of assessment; standardised assessments that are out of date (older than 10 years or 15 years for cognitive tests) and assessments that were never standardised in the first place.

A good example of a non-standardised assessment is a Miscue

[4] Information on assessments mentioned in the paper can be found in appendix 7

Analysis. This is a technique in which a pupil's strategies are assessed when he/she is reading. The idea is to document how the pupil attempts to read all the words in a given text, these errors are recorded as substitutions (words read incorrectly), self corrections, refusals, insertions and omissions (Backhouse, 2005, p173). The value of the documented errors from a Miscue Analysis can be seen from the following example;

A pupil is reading a text for the Miscue Analysis and pauses on a word, the strategy he/she employs to read that word (and the effectiveness of that method) may give clues to their (proximal) difficulty with reading. For example, the pupil may sound out the letters of the word and make errors doing this. This information may lead you to carry out further assessments using the standardised Comprehensive Test Of Phonological Processing (CTOPP, PRO-ED, 1999).
Crucially, the diagnostic skills gained by carrying out non-standardised assessments can be employed when using other tests or when analysing samples of their work. An example of this technique can be seen when using the Alpha to Omega placement test or a list of high frequency words. Because these are not standardised, the results shouldn't be used when reporting the outcome of assessments, however, by analysing the errors, and the pattern of these errors in the list of spellings difficulties including initial sounds, onset and rhyme, letter translation/rotation can be diagnosed. This same technique of spelling analysis can be used on samples of the pupils work and carried over to use on other assessments, for example, the spelling subtest of the Wide Range Achievement Test forth edition (WRAT 4).

Standardised Assessments.

A standardised test is one in which it is administered in a standard way. A mathematics test can be administered following a set number of rules or conventions; for example, pupils cannot use a calculator or other such equipment and must stop the test once they have exceeded the time limit. It is obvious here that a pupil who sat the same test but used a calculator and had an indefinite time limit would probably score better than his or her peers. By doing this, the test has been invalidated, as it hasn't been sat within the required conditions. The standardised psychometric tests fall under the Norm referenced category, that is not to say that they can't be used in an ipsative way – when comparing the progress over a period of time using the same test. A standardised assessment is often a useful way to confirm the teacher's views/concerns regarding a pupil, an objective standard against which his/her ability can be measured. This is in contrast to an informal assessment where the results cannot be compared due to differences in application.

Standardisation Techniques.

A discussion of standardisation techniques will include mention of several technical terms. Terms in green will be looked at in detail later but to gain a general idea of a standardised test it will be useful to construct one in theory.

The first step in constructing a standardised test is to decide what proximate difficulties the assessment will focus on and to make sure the questions actually measure this difficulty, this is referred to as the validity of the test. These questions then must be tested to see if they accurately relate to the proximate difficulty and that work consistently, this is referred to as the reliability. Once a set of questions have been decided upon the test is given to number of people known as the test population, N. The results of these tests are then analysed statistically[5] to produce a distribution curve so that subsequent test results – often given as standardised scores (a score that incorporates an individuals test score with their age) can be plotted against this curve to give a norm referenced score.

Distribution curve.

The distribution curve (or bell curve) is a graphical representation of a sample population of random variables that tend to cluster around a mean value. What this means for assessment data is that the results of the test/sample population (N) are plotted around the average (mean) result for the test. Once the test is completed for an individual, they can then be compared against the population. A full description of the construction of the distribution curve can be found in appendix 2, a history of it's origins can be found in appendix 4.

Among the most confusing terms used is that of randomness. When randomness has been used so far it relates to a 'closed' type of randomness, for example, the Galton pinball machine is random but will only give results marked in the six collection trays from the image – there is no chance of the pinball ending up in the seventh, or even the thirty-seventh tray as this is not possible. In this sense, randomness should be taken to mean the randomness experienced while gambling[6] (one of the horses will win the race – not a horse that wasn't there at the start, this is randomness in a closed system. It is with particular view of a closed system that psychometric testing can be placed in. Take IQ for example, there is an expected range (closed system) in which the results will fall – it is very unlikely, or even

3. A full description of this technique is given in appendix 2.
Gambling is very much a closed system – random events do occur within it, but it is demarcated from true randomness in so much as a player can never have five aces.

measurable, that someone should get an IQ of seven hundred. Therefore IQ can be treated as a closed system.

The bell curve may have origins unrelated to assessment but it is now used in every standardised test (norm referenced) the strength (or weakness) of the technique of using a distribution curve relies on the size of N, or the size of the sample population.

Population size of the sample (N).

The strength, or weakness of the Bell Curve is dependent on N – the size of the population tested. Extreme examples of this can be found from secondary school mathematics when discussing the probability of coin tosses. The meaningless phrase is often used; "an infinite number of coin tosses will result in a 50/50 out come of heads and tails". Here the N is infinite and therefore meaningless, but if we look at a small sample (smaller N) of say, ten tosses, it is quite likely that we may see eight heads and only two tails. If this is the final sample then we can say that the probability of getting a tail is 20% or two in ten, this is obviously mistaken – but only because we know that the more tosses will result in an outcome nearer that of 50%. This may seem obvious but this confusion of N (deliberate or otherwise) is quite widespread. Take one statistic; one in ten children have working memory problems (Packiam, Alloway, 2008, PATOSS Conference). We assume (hope) that this was taken from a large sample population (N). The problem arises when this large population (N) is then used to refer to a significantly smaller population (N1) 'therefore out of a class of thirty, you will have three children with a problem with their working memory'. At this time it is very tempting to think of children that may have this problem, of course, because you have just been told they are there.
This is a problem, however, involving the size of the population (N) - consider a similar statistic; just over 30% of the UK population has a degree.
(http://www.education.gov.uk/rsgateway/DB/SFR/s000798/DIUSSFR05-2008.pdf)
If it was then suggested that a smaller population (for example the attendees at the OCR course where everyone has a degree) followed the same trend then it would not be seen to fit and would be dismissed. The analogy with the coin tosses (N = 10) can be seen here, the result of 20% chance of tails is known to be wrong because, and only because we know better, when it comes to receptive language (as measured by the BPVS III) we don't know the result until the sample is taken.

From this brief look at N, the following issues have been identified;

- The effect of a large closed population (system) when used on a

smaller closed system – the national statistic of degree holders vs the attendees of the OCR course and the problem of children (nationally) having problems with working memory vs the small closed system of a class of thirty children.

- The opposite effect of a small population when used on larger systems – the N = 10 coin tosses to give a prediction of 20% tails events occurring.

A breakdown of N for both the BPVS and the CTOPP can be found in appendix 3 where a particular age (6) was chosen to highlight the number of individuals of that same age appearing in the sample – who they would be compared with (less than 200 in each).

Reliability.

Reliability is the consistency of the set of measures of a test. It is the ability of a test (or a better example would be the two parallel tests found in the WRAT test) to produce consistent results. Pupil A is tested first with the blue form and then with the green form from the WRAT, the reliability of the test would be measured upon the similarity of the results obtained from the same pupil. If one were to use a cold drinks machine and pressed 'Coke' three times and you received a 'Fanta', a 'Sprite' and a 'Diet Coke' then the machine would be labelled unreliable. An important distinction between reliability and validity can be made here with this analogy, if the 'Coke' button were pressed three times and you received three cans of 'Fanta' the machine would be said to be reliable but not valid.
A perfectly reliable test would give a reliability coefficient (Cronbach's alpha, to be referred to as simply alpha) of 1.00.

Inter-rater reliability – this is the variation of results obtained by different people taking the same test.

Test-retest reliability – this is the variation of results obtained by the same person taking the same test.

Internal consistency reliability – this is the consistency of results obtained from the questions in the test.

Internal consistency is a measure of the correlations between questions within a test. "Correlation is a way of describing how closely two things relate to each other. It is expressed as a number called a correlation coefficient. The range of values that a correlation coefficient can take is from -1 through to 1 [closer to 1 being more consistent]" (Deary, 2001). For example, the math section of the WRAT measures mathematical computation, and the collection of questions (with

increased complexity) are mathematical computation questions. These questions would correlate together as they are all mathematical computation questions measuring mathematical computation. If there was a question which required the pupil to interpret a graph this would be measuring the pupils data handling and mathematical interpretation skills, not mathematical computation. A question such as this would not correlate with the other questions in the section as it is measuring a different aspect of mathematics.

Reliability does not imply validity (see below), a test may be reliable but may not measure what you want to be measuring e.g using the BPVS III for an intelligence test.

The CTOPP breaks down each subtest (CTOPP, p.69) and found that all of the subtests attain an alpha of 0.70, 76% attain 0.80 and 19% attain 0.90. The alpha for the composite scores exceed 0.80. Where; Strong correlation = 0.5 – 10, Medium = 0.3 – 0.5, Small = 0.1 – 0.3 and None = 0.0 – 0.09 (Deary, ibid).

Validity.

Validity can be seen as the degree in which the test measures what it claims to measure, validity can be split into three types;

Construct validity – this is the theoretical and empirical evidence supports the claim of the a test to actually measure what it says it does, for example how the CTOPP (Comprehensive Test of Phonological Awareness) test measures phonological awareness, phonological memory and rapid naming (speed of processing).

Content validity – this refers to the extent to which the content of the test measures the range of behaviours expected from the theory, for example from the CTOPP, the content validity of the phonological awareness portion of the test would be measured on the test questions addressing a wide range of phonological awareness difficulties – do the questions/tasks measure detection and manipulation of syllables, onset and rhyme and phonemes? The CTOPP deals with the content validity question by showing that the abilities chosen to be measured are consistent with the current knowledge about a particular area e.g. phonological awareness.

Criterion validity – this is the ability of the test to produce evidence that is comparable with other tests that measure the same difficulty. For example, you would expect a pupil obtaining a poor comprehension score produced by the NARA (Neale analysis of reading ability) to also

acquire a poor comprehension score from the Access Reading Test (wide-range reading assessment).

The BPVS III makes it clear that the test is designed for children that follow the two conditions; when English is the language of the home and community in which the pupil has grown up and resides in, and when English is, and has been, the primary language of instruction at school – making it unsuitable for EAL (English as an Additional Language) children.

Using a wide range of educational and psychometric assessments.

As discussed in the introduction, an ultimate difficulty can be caused by a number of different proximate difficulties (and different combinations of them). From this, one can see that a specialist teacher would require a wide range of assessments in order to discover what proximate difficulties a pupil has.

Ultimate difficulty measured	Assessment	Proximate difficulties
Reading comprehension.	Access Reading Test	Literal Understanding of vocabulary. Comprehension requiring inference or prediction and opinions. Comprehension requiring analysis
	NARA	Reading accuracy. Reading Comprehension. Rate of reading.
	Salford Sentence Reading Test	Reading (sentence).
	Vernon Warden	Reading (sentence).

From the table it is clear that the ultimate difficulty of reading is broken down into different proximate difficulties by these four different tests.

Any combination of these proximate difficulties may be the cause of the ultimate difficulty of reading.

It is important to note that while a subtest is designed to produce a score that will inform a diagnosis (remembering the rules of implication), a subtest may give a score that sheds light onto a proximate difficulty, but practice of that subtest will not improve what the assessment sets out to test. A number sequence in reverse may give concerns relating to a difficulty with working memory, but practising recalling similar number sequences will not improve working memory.

Careful observation of the student while assessing by the specialist teacher can also provide invaluable information towards diagnosis. How are they sitting? How do they react to each test? The assessor would also take note of any intrinsic (not feeling well) and extrinsic (environmental/external) factors that may affect the results of an assessment.

Selection of tests.

A number of tests have been mentioned during this discussion of psychometric and educational testing, a question that specialist teachers would need to ask themselves is – what tests do I use?

Test selection can be narrowed down by two fundamental variables; the perceived problem (ultimate) – this is supplied by the background information on the pupil, questionnaires etc (Ott, 1997), and the age of the pupil (what phase of education they are in). For example, a pupil (6) may have been identified with reading (comprehension) difficulties. Both the Access Reading test and the NARA measure reading comprehension, however the NARA has an age range of between 6 and 12:11 whereas the Access has an age range of 7 to 20+, therefore the Access would not be suitable. The time available for testing is also a consideration when selecting tests and the specialist teacher may wish to use just the subtests of standardised tests in a non standardised diagnostic manner.

An important consideration is the intention or aim of the assessments. Assessments can be used to uncover a specific learning difficulty or to qualify an individual for exam access arrangements (extra time, scribe, reader etc.). For exam access arrangement the selection of tests are stipulated by the JCQ (Joint Council for Qualifications), where qualifying assessments will be listed.

Finally the specialist teacher needs to be aware of confidentiality and insurance. Every assessment that they perform has to remain

confidential because of the connotations of labelling students, the aim of assessment is not to label pupils but to use the information collected to inform and implement a teaching programme to help the student. Insurance is essential[7], where a diagnosis has the possibilities to affect the life (educationally, emotionally and socially) of an individual. Once a diagnosis/report has been made the specialist teacher cannot control what is done with the information (how recommendations are implemented), or how it is treated, and this can have serious implications for the students throughout their lives.

Conclusion.

An important distinction has been suggested by this discussion; that of ultimate and proximate difficulties. Without a clear understanding of the proximate difficulties and how they can be identified and improved upon we are left with the misconception see so frequently in schools where an ultimate problem is identified (reading) by the proximate difficulty comprehension (via the non-statutory yearly QCA test) and then the pupil(s) are put into an intervention group (e.g P.A.T) to improve their proximate difficulty of phonological awareness. Psychological and educational assessments are importantly seen as a tool for the specialist teacher to use, in order to investigate the more simplistic (but more universally understood) ultimate difficulties in order to highlight the proximate difficulties; and more importantly the combination of these that will lead to a diagnosis of a specific learning difficulty and a carefully planned programme of support and intervention that will help the pupil work to improve or compensate for their difficulties and enable them to succeed.

- § -

[7] A specialist teacher working in a school or LEA is covered by the institutions insurance. Someone working independently can obtain insurance by joining a professional body such as PATOSS.

Appendix 1

Conditionals.

$$a \supset c$$

(if antecedent then conditional) (Quine,1950, p.21)

The conditional ⊃ , should be read as if__then, a few examples will help here;

1) If I trip over the toy then I will fall down the stairs.

2) If the fuse is broken then the washing machine will not work.

Statements 1 and 2 can be rewritten to include the conditional;

1) I trip over the toy ⊃ I will fall down the stairs.

2) The fuse is broken ⊃ The washing machine does not work.

These examples show how the antecedent is responsible for the conditional. By rearranging examples (1) and (2) it can be demonstrated that although $a \supset c$ maybe true it does not logically imply that $c \supset a$ is also true.

a) If I fall down the stairs then I have tripped over the toy.

Notice here that this statement might be true but many other factors may have caused me to fall down the stairs, just because (1) was true it does not mean that (a) must be true, similarly;

b) If the washing machine does not work then the fuse is broken.

This example makes the case clearly. Statement (2) is true as a broken or faulty fuse will stop the washing machine working. It is obvious that just because (2) is true it doesn't make (b) necessarily true, otherwise washing machine mechanics would have an easy job just replacing fuses as statement (b) clearly states that this is the reason for washing machines to break.

Appendix 2

Creating the bell curve.

Twenty valid and reliable questions have been collected (therefore giving a range of 0 – 20 in possible answers). The test is then given to N, the sample population e.g 100 6 year olds8, here are some results;

	Score
Candidate 1	17
Candidate 2	12
Candidate 3	7
Candidate 4	15

Once all the scores have been collected the sample mean average can be calculated;

Sum of scores ÷ N = sample mean

Next the deviation from the mean (we'll say it's 15) is calculated for each candidates score. Because some candidates will score below the sample mean (which will result in a negative score) the deviations are squared.

	Score	Deviation from the mean (-) sample mean (15)	(squared) Deviation
Candidate 1	17	2	4
Candidate 2	12	-3	9
Candidate 3	7	-8	64
Candidate 4	15	0	0

From the table you can see that candidate 4, who scored the average

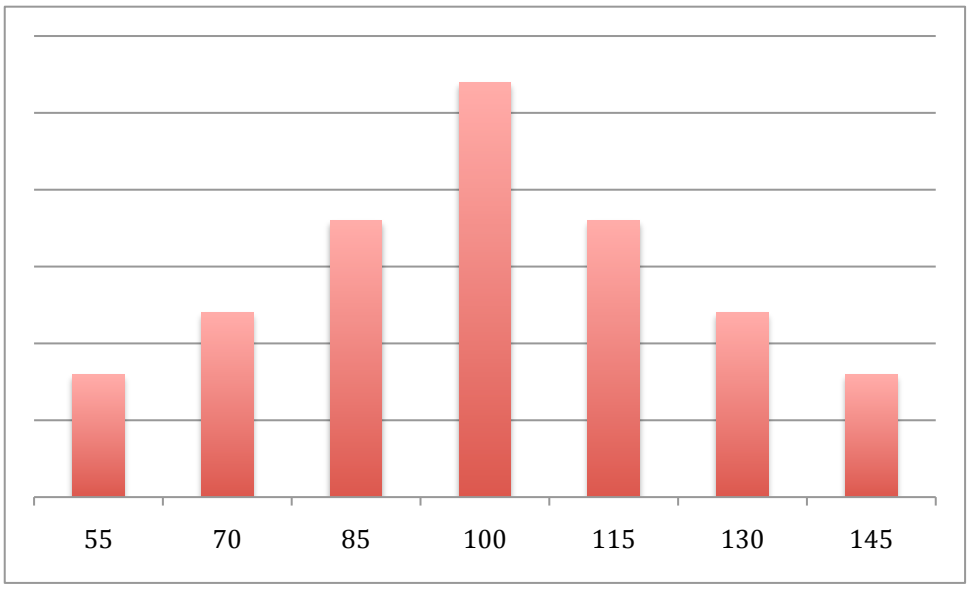
To avoid the problem alluded to in the discussion of N, the CTOPP made sure that each age group contained a representative sample from the following groups; geographical area, gender, race, residence (urban, rural), income (family) and parental education.

score, deviates from the mean by 0.

The standard deviation is calculated by adding up these deviations and dividing the total by the sample population minus 1.

Sum of deviations ÷ (N – 1) = standard deviation.

The bell curve is plotted by placing the sample mean at x = 100. For simplicity, let's say that the standard deviation is 15. The next plots on the graph would be +/- the standard deviation, so lines would be marked at 85 and 115, then at 70 and 130 and finally at 55 and 145.



(This chart is meant to demonstrate the positioning of the points along the x axis, the peaks on a standard bell chart would be connected together in a smooth continuous line.)

The height of the y axis at the mean (100) is unimportant, but the subsequent plots fall by a third.

BPVS III

Standardisation took place in February 2009, with a total of 3278 students from 147 schools taking part (with one year trialling in 14 schools – total of 161 schools).

Looking at a particular pupil (6 years old) in the 6-7 year old category, you can see that their test score will be positioned against 190 students scores.

Age/year group				No. of students	No. of schools
Age	**England and Wales**	**Scotland**	**Northern Ireland**		
3 - 4 years	Nursery/pre-school	Nursery/pre-school	Nursery/pre-school	404	27
4 – 5 years	Reception	P1	P1	225	17
5 – 6 years	Year 1	P2	P2	200	15
6 – 7 years	Year 2	P3	P3	190	14
7 – 8 years	Year 3	P4	P4	188	14
9 – 10 years	Year 5	P6	P6	423	17
10 – 11 years	Year 6	P7	P7	382	15
11 – 12 years	Year 7	S1	Year 8	414	17
12 – 13 years	Year 8	S2	Year 9	364	12
14 – 15 years	Year 10	S3	Year 10	305	8
14 – 15 years	Year 11	S5	Year 12	183	5
Totals				3278	161

(BPVS III, p.28)

CTOPP

Standardisation took place between autumn 1997 and spring 1998 with a total of 1656 persons in thirty states (in America).

Looking at a particular pupil (6 years old) in the 6 year old category, you can see that their test score will be positioned against 155 students scores.

Demographic Characteristics of the Normative Sample	
Age	**No. of Students**
5	149
6	155
7	140
8	151
9	155
10	126
11	131
12	108
13	106
14	93
15	76
16	77
17	77
18 - 24	112

(CTOPP, p.61)

Appendix 4

"The ubiquity of the Gaussian is not a property of the world, but a problem of our minds, stemming from the way we look at it" (Taleb, 2004)

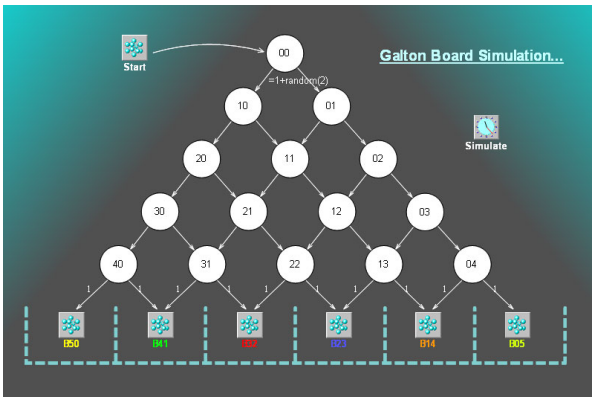
In probability theory, the Bell curve (Normal or Gaussian distribution) is often used for random variables that tend to cluster around a single mean value. Appendix 2 provides an explanation on how the bell curve is produced.

Abraham de Moivre (1665 – 1754) who wrote a book on probability theory, 'The Doctrine of Chances', made the first recorded discovery of the (then termed) curve of error in 1733 while investigating games of chance.

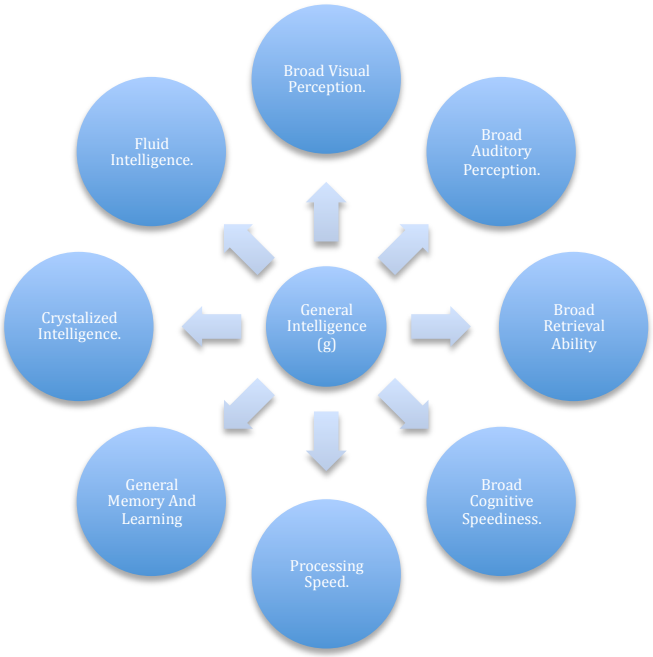
Carl Friedrich Gauss (1777 – 1855), a german mathematician, noted that most observations (events) hover around the average, and that, the odds of deviation decline exponentially as you move away from the average (the Bell curve). Because of this fast deviation (steepness of the curve away from the mean) there is a vulnerability in the estimation of tail events – particularly of interest to those looking at this area of the curve (the low ability). The Bell curve, known as the Gaussian was initially meant to measure astronomical errors.

Adolphe Quételet (1796 – 1874) came up with the notion of a physically average human – height, weight etc. But quickly moved on to social matters (Taleb, 2007). This led on from his construction of *l'homme moyen maral* and *l'homme moyen physique* (morally and physically average man) to deviations becoming abnormal. This sense of the abnormal also influenced contemporary thinkers, "Social Deviations in terms of the distribution of wealth for example, must be minimised" (Marx, 1867). This thinking became so ingrained during Quételet's time that the Bell Curve was referred to as *le loi des erreurs* – the law of errors, 'ought' is confused with 'is'. More importantly, for our discussion, was that; "Quételet's doctrine, like Aristotle, exempted mental abilities, arguing that those superior to the average in intelligence were mere forerunners of a new average that was to come". (Fashing, 1981)

Sir Francis Galton (1822 – 1911) rediscovered the technique when he constructed pinball machines;



A hierarchical representation of the associations among mental ability test scores. This diagram was the result of decades of work by John B. Carroll who re-analysed over 400 large, classic databases on human intelligence research.



(p.14, Deary, 2001)

Appendix 6

The Wide Range Intelligence Test (WRIT)

General Intelligence

Verbal (Crystallised) IQ	Verbal Analogies	Vocabulary
Visual (fluid) IQ	Matrices	Diamonds

WRIT manual p.2 adapted from figure 1.1

Similar to the Wechsler, the WRIT contains different task that have a positive correlation with each other.

Appendix 7

Access Reading Test (wide-range reading assessment)

Publisher	Hodder Murray (Hodder and Stoughton) 338 Euston Road London NW1 3BH Tel: 020 7873 6000 http://www.hoddereducation.co.uk/Home.aspx
Date of publication	2006
Cost	Set (manual and one copy of Forms A & B) £25.00 Set of 10 copies of Forms (A or B) £12.50 Interactive CD Rom – single user £125 Network £400
Focus of assessment attainment;	Assessment of pupils reading • literal comprehension. • Understanding of vocabulary. • Comprehension requiring inference or predicion and opinions. • Comprehension requiring analysis.
Standardised/non standardised	Standardised (4092) • higher average = 110 - 115 • average = 85 - 115 • lower average = 85 - 90
Age range	7 - 20+
Group/individual	Both
Comments	• 30 minute time limit.

- Form A or B
- Pupils who have not completed the test in 30 minutes may then be given 25% extra time (7 ½ minutes) if you feel that this is appropriate and you wish to see if there is an improvement in their score with the extra time. They should mark on the paper where they reached after 30 minutes by writing '30 mins' against the question number, and change to a different colour pen.
- Suitable for access arrangements

BPVS III (British Picture Vocabulary Scale)

Publisher	GL assessment http://www.gl-assessment.co.uk/
Date of publication	2009
Cost	Set - £157.50 Record forms - £10.75
Focus of assessment	Receptive language
Standardised/non standardised	Standardised. (3278)
Age range	3 - 16
Group/individual	Individual.
Comments Ravens	Pupils with large discrepancy between and BPVS should be looked at further. A high BPVS and low Ravens may indicate a pupil with dyspraxic type difficulties. A referral to Occupational Therapy via school nurse or doctor is advisable.

A high Ravens and low BPVS may indicate speech and language problems, and a referral to the mainstream Speech and Language Therapy service should be considered.

(Lewisham SEN profile checklist 2005)

Takes around 10 minutes.

CTOPP (Comprehensive Test of Phonological Processing)

Publisher	PRO-ED. http://www.psychcorp.co.uk/
Date of publication	1999
Cost	Full kit £274.50 (plus VAT) – includes: Manual, Audio CD, Picture Book, Record Booklets Additional Pkts of 25 Record Booklets: Ages 5/6 £61.20 plus VAT Ages 7-24 £72.00 plus VAT
Focus of assessment	• Phonological awareness. • Phonological Memory. • Rapid naming (speed of processing).

Standardised/non standardised a stratified	Standardised. (1997 - normed on sample of 1,656 individuals)
Age range	5 - 24
Group/individual	Individual.
Comments	• Poor performance on rapid naming tasks tends to result in reading fluency. • Additional subtests; • Phoneme reversal – reordering speech sounds to form words. • Segmenting words – say the separate phonemes that make up a word. • Rapid colour naming and rapid object naming – are further rapid naming tests assessing processing speed. Can support evidence for extra time if tests indicate o Cognitive processing difficulties (rapid naming tests) o Phonological memory and awareness difficulties (word manipulation and word blending tests) • Takes time but very informative Salford Sentence Reading Test
Publisher	Hodder Murray (Hodder and Stoughton) http://www.hoddertests.co.uk/
Date of publication	2002
Cost	£24.50

Focus of assessment	Reading (sentence)
Standardised/non standardised	Percentile rankings based on original norms set the score in context of 100 pupils.
Age range	6 - 10+
Group/individual	Both
Comments	• Two parallel tests X and Y • Stop after 6th error. Complete the sentence, allow self-correction. Supply word after 6 seconds. • Reading age is at the 6th error, Reading ages below 6 should be viewed as 'statistical reading ages' only.

Vernon Warden

Publisher Available at www.dyslexiaaction.org.uk

Focus of assessment	Sentence reading (multiple choice).
Standardised/non standardised	Standardised (British standardisation 1760)
Age range	8 - adult
Group/individual	Both.
Comments	• can be considered a measure of reading accuracy, fluency and comprehension. • No guessing. • re-standardised in 1993 and 1994 in Kirklees, UK • Up to and including year 8 – 15 minutes. • Year 9 and above – 10 minutes (unless extra time is an issue)

WRAT 4 (Wide Range Assessment Test)

Publisher	Pyschological Assessment Resources Inc. (USA) Ann Arbour http://www.annarbor.co.uk
Date of publication	2006
Cost	Introductory kit £194 (Exc VAT) Sets of additional test forms and/or
Response	Booklets (25) £29.00 exc VAT
Focus of assessment	• Sentence Comprehension • Word Reading • Spelling • Math Computation
Standardised/non standardised	Standardised (+3000)
Age range	5 - 94
Group/individual	Individual Both (Spelling and maths in small groups).
Comments	• Two parallel tests Green and Blue • Assesses single letter name knowledge. • 15 – 25 minutes for ages 5 - 7. • 35 – 45 minutes for ages +8. • suitable for exam access arrangements.

- Some difficulty can be caused with Americanisms.

WRIT (Wide range intelligence test)

Publisher	PAR (Psychological Assessment Resources, Inc).
	http://www.hogrefe.co.uk/
Date of publication	1999
Cost	Kit - £210 Additional forms - £38
Focus of assessment	Verbal-Crystallized Abilities • Vocabulary (a traditional word definition task). • Verbal Analogies (a fast-paced verbal-reasoning task). Nonverbal-Fluid Abilities • Visual Matrices (a traditional matrix task assessing visual-spatial reasoning and abstract visual-perceptual relationships) • Diamonds (a spatial constructional task using diamond-shaped chips to duplicate two- and three-dimensional figures)
Standardised/non standardised	Standardised. (2285)
Age range	4 - 85

Group/individual	Individual.
Comments	• helps to document ability levels and cognitive ability • helps to identify learning disabilities, giftedness, neuropsychological impairments and other exceptionalities. • When used with the Wide Range Achievement Test (WRAT) with which it is co-normed, allows for sound and efficient identification of intelligence/achievement discrepancy • Should take between 20 – 30 minutes • American accent may confuse some students.

References.

Backhouse, G & Morris, K, 2005 **Dyslexia? Assessing and reporting**. Bristol, Hodder Murray

Deary, I.J, 2001 **Intelligence**. Oxford, Oxford University Press

Dunn, L.M et al **BPVS III examiner's manual**

GL assessment

Education.com Available at: **http://www.education.com/definition/receptive-vocabulary/** [Accessed 12/01/11]

Education.gov.uk Available at:

http://www.education.gov.uk/rsgateway/DB/SFR/s000798/DIUSSFR05-2008.pdf

[Accessed 14/01/11]

Fashing, G, 1981. Available at: **http://crab.rutgers.edu/~goertzel/normalcurve.htm**

[Accessed 12/01/11]

Ott, P, 1997 **How to Detect and Manage Dyslexia**. Oxford, Heinemann

Quine, W.V, 1950. **Methods Of Logic (4th ed)**

Cambridge (Mas), Harvard University Press

Taleb, N N, 2004. **Fooled By Randomness**. London, Penguin

Taleb, N N, 2007. **The Black Swan**. London, Penguin

Waine, L & Kime, S **Practical Aspects of Assessment** In: **Dyslexia? Assessing and reporting** Bristol, Hodder Murray

Wagner et al, 1999. **CTOPP examiner's manual** Austin, PRO-ED

Bibliography.

Backhouse, G & Morris, K, 2005 **Dyslexia? Assessing and reporting**. Bristol, Hodder Murray

Beard, R, 1990 **Developing Reading 3-13 (2nd Edition)**

London, Hodder & Stoughton

Booth, T et al (Ed.), 1987, **Preventing Difficulties in Learning** Oxford, Blackwell

Deary, I.J, 2001 **Intelligence**. Oxford, University Press

Dunn, L.M et al **BPVS III examiner's manual** GL assessment

Education.com Available at:
http://www.education.com/definition/receptive-vocabulary/
[Accessed 12/01/11]

Education.gov.uk Available at:

http://www.education.gov.uk/rsgateway/DB/SFR/s000798/DIUSSFR05-2008.pdf

[Accessed 14/01/11]

Fashing, G, 1981. Available at:
http://crab.rutgers.edu/~goertzel/normalcurve.htm

[Accessed 12/01/11]

Fawcett (Ed.),2002**Dyslexia, Theory & Good Practice**
London, Whurr

Marsh, J & Hallet, (Ed.), 1999, **Desirable Literacies**
London, Paul Chapman

Ott, P, 1997 **How to Detect and Manage Dyslexia**.
Oxford, Heinemann

Quine, W.V, 1950. **Methods Of Logic (4th ed)**
Cambridge (Mas), Harvard University Press

Reid, G. (Ed), 1996 **Specific Learning Difficulties (Dyslexia)**
Glasgow, Moray House

Snowling & Thomson (Ed.),1991 **Dyslexia, Integrating theory & Practice**
London, Whurr

Taleb, N N, 2004. **Fooled By Randomness**.
London, Penguin

Taleb, N N, 2007. **The Black Swan**.

London, Penguin

Waine, L & Kime, S **Practical Aspects of Assessment**

In: **Dyslexia? Assessing and reporting**
Bristol, Hodder Murray

Wagner et al, 1999. **CTOPP examiner's manual**
Austin, PRO-ED