Assessment:

The psychological science behind assessments (and some lies)

Education, is sometimes parsed into three core activities, curriculum, pedagogy and assessment. Regardless of how we believe children or students show intelligence, and therefore in part some of what we are trying to teach them (in whatever arena or topic area), at some point we need to figure out whether this teaching has been effective, or not. If it is, then all well and good. If it has not, then either we've been poor teachers, or the student was not attending, or the student could not understand: or it's a combination of any of these issues.

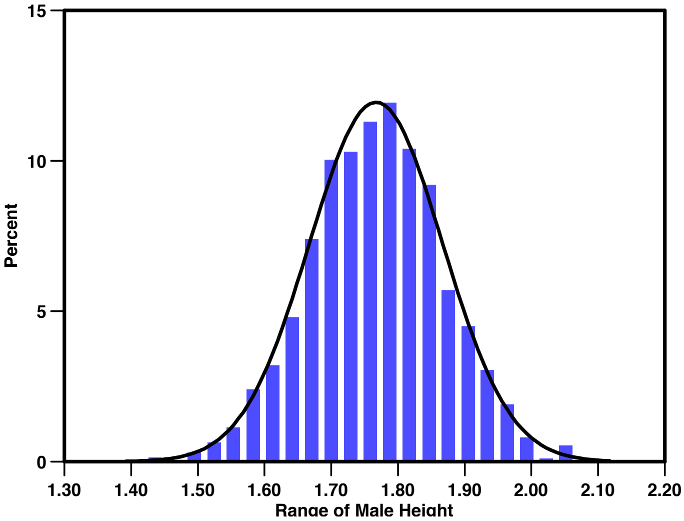
This lecture is about the psychological science, and in some cases the apparent science but which is not actually supported by psychological knowledge, in the way that assessment is constructed. This lecture is not supposed to replace the assessment issues that will be covered later in your course, but merely to put you on track as to the psychological theory (or not) behind why some assessment is utilised.

Somewhat shocking (to me) when I first encountered this was that assessment is probably as much about a philosophy of what it means to be educated, as it is about the science that informs assessment policy.

A Scientific Approach to Educational Assessment

This is the classic 'text book' explanation of how assessment in education is constructed. It looks and sounds very scientific.

The apparent frequency distribution of human abilities



I apologise but this should be the last bit of 'statistics' that I give you for this course. Namely the 'bell curve' distribution that apparently is supposed to occur in most human abilities. Take human height for instance. We find that most people clump around a middle score, a few folks deviate away from this middle score, and a very small number deviate at extreme heights. If you look at the graph to the right, you see the frequency distribution of 2000 males from the USA randomly sampled and their height measured. A curve has been fitted over the top of the actual bars that have been plotted, so you can clearly see the classic 'bell shape'. If you've ever heard anyone talk about the *standard deviation* this simply indicates the average distance away from the arithmetic mean. The arithmetic mean is the top of the 'bell'. If you go either side of this peak and mark off about half way down the steep slope, that's about the average distance away from the arithmetic mean. This value from the arithmetic mean, to the average distance away from it, is called the *standard deviation*. When you think about the term '*standard deviation*' it kind of makes sense no?

OK, now a distribution like this is called a 'normal distribution' (sometimes it's called a Poisson distribution). It has the following properties which I'm not here to prove, but simply to tell you: 68% of a population that is 'normally distributed' falls within 1 standard deviation either side of the arithmetic mean. 95% of the population falls within 2 standard deviations either side of the arithmetic mean, and finally 99% of the population falls within 3 standard deviations of the arithmetic mean. That's pretty much it.

Mathematically there's nothing particularly special about this distribution, but in social sciences it plays a huge role, because it's assumed that most things that social scientists are interested in measuring falls in this distribution.

OK, now I hope that this is all that you need to understand the rest of the lecture.

Norm Referenced Assessment

Social scientists started to notice the normal distribution turning up everywhere one measured anything in human society (and even in animal and plant populations). Weight, head circumference, vocabulary memory, reaction time and so on. It started to become less of a surprise and more of an accepted distribution that applied to living organisms, including humans.

So the reasoning went, we assume that intelligence is also distributed normally (see the circular argument box). The results has been that educators have tried to use assessment methods that specifically reflect that normal distribution, with the claim that it has good 'discriminatory power'. A fancy way of saying that it shows us what we were expecting to see. If you're results do not exactly show a normal distribution (it might be 'skewed') then that's ok too. Since these educators 'know' that the distribution of ability is normally distributed (since the psychologists told them so!), it's quite easy to make mathematical corrections to the assessment scores, so that the top 4% (or so) of the students achieve an 'A' grade, the next 20% get a 'B' grade, the following 40% get a 'C' grade and hey one can split the remainder into 'D' and 'E' and call it even. What this means of course is that every year will show the same distribution of grades, regardless of how well the pupils actually did. I've been in classes where the teacher was able to mark maths papers for national exams before they were handed in and where the whole class did really well, getting 'A' or 'B' grades compared to previous years. Yet when the scores were officially returned they were all back to 'B's and 'C's. The teacher shrugs and tells us that unfortunately the paper was a bit easier and therefore the national assessors made a correction to force our scores back into a normal distribution.

Norm referenced circular arguments

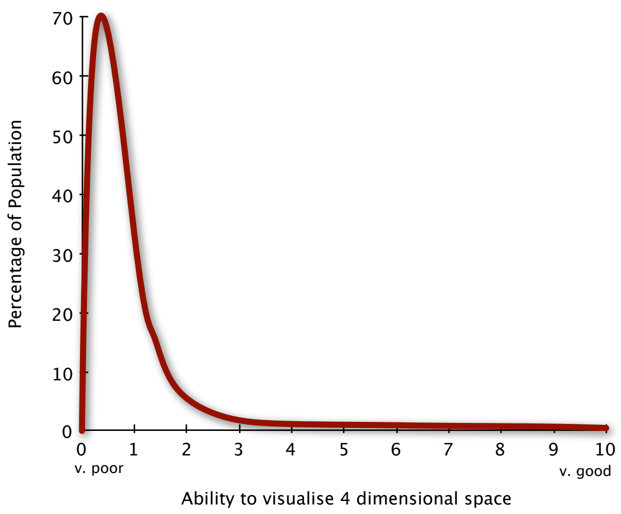
I'm always stunned by the arrogance of people from psychometrics when they are (I'm not trying to say that all psychometricians are arrogant) trying to prove that intelligence is normally distributed throughout humanity.

Since social scientists had noticed the normal distribution in some many arenas of humanity, it seemed 'obvious' that mental intelligence would also follow a normal distribution.

Since it was 'known' that intelligence was normally distributed, psychometricians designed tests whose scores were indeed normally distributed. They did this by keeping in test items that showed the normal distribution and discarding the rest.

One could still live with this state of affairs except that well meaning psychologists have stated and taught countless people that intelligence is normally distributed as shown by the results of the intelligence tests.

Err hello…????



If I sound cynical about this – I am; **not** because the experience has left me a bitter and twisted person (my skin is thick enough to get over these kinds of incidents), but because there is a sense of academic dishonesty about this. Firstly, there is the dubious 'science' as to whether intelligence is really normally distributed. I'm not saying it is not, but I'm not saying it is. We do not have (i) a good definition of what intelligence is to be able to measure it correction and (ii) we do not have objective scores that can measure such an intelligence in a satisfactorily objective and culture free way. How do we know that children's intelligence is not actually distributed in a very bimodal manner (you either 'get it' or you 'don't'), or something like most people are clumped towards one end of a spectrum and only a few outliers are clumped at the other end. Take for instance one's ability to see into 4 dimensional space (a 3D cube folded out into 4D is called a *tesseract*). Not many people can actually 'see' this – you can intellectually understand it – but visualising it is a different matter. If I was to draw a frequency distribution graph in the population that can actually visualise a tesseract it might look like the attached graph. What you notice is that most of us cannot see the whole tesseract (essentially because our vision has been developed for three dimensions, not four), but a very small number of us can. Is there any reason to believe that our intelligence is not actually distributed in this way? Or another way (the reverse of this), or something we've never thought of or seen?

Perhaps more damagingly, how does dividing the school child population into discrete cohorts ('A', 'B', and 'C' graders etc.)? One might argue that these assessments allow us as educators, or as employers to select the 'best' in a cohort of children.

But do we?

There are no strong, or consistent correlations between IQ scores and job performance. The American Psychological Association, made a general statement that suggests essentially that IQ may contribute somewhat to job performance, but other measures, such as interpersonal skills and other aspects of personality are 'probably of equal or greater importance'. The same report in the next paragraph suggests that higher IQ scores are also correlated more with unsocial behaviour such as juvenile crime.

In this section, what I've tried to do, is to show you the 'apparent' science behind educational assessment that has been going for many decades. In part it was inspired by the psychological research done into intelligence testing and it's apparent distribution throughout the general population (we've dealt at length with this issue over the previous two lectures).

This begs the question what kind of assessment should we employ then?

Criterion Referenced Assessment

The answer is '*criterion' referenced assessment*'. That is we don't consider where the relative position of a child is compared to her or his classmates, but rather whether they can actually do a task or not. Sometimes this is called 'performance based' assessment, or 'task competency' or 'benchmarking' but essentially they are the same thing.

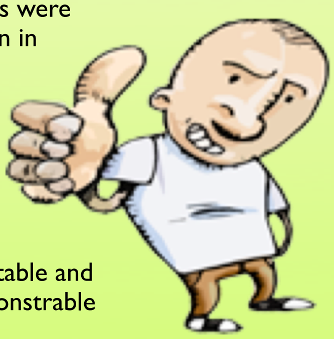
The logic is this, state what you want the student/pupil to be able to do if they have successfully learned the thing that you're teaching. Furthermore, what they should be able to do must be measurable, it cannot be 'a feeling'. It has to be something that can be stated and reproduced by someone else. If you do this, then someone who is thinking about working with your student/pupil knows what task they can do.

The case for Criterion Referenced Assessments

Why, some of you may ask, should we not employ a norm referenced approach. Who would not want to employ the best people from an age cohort? The answer is this, your relative position does **not** necessarily tell anyone what you're actually capable of.

You're holding the hand of your child (or brother or sister if you're not a parent yet) as they lie on the operating table just before the big open brain surgery that's about to be performed on him or her – it's a necessary but currently not life threatening operation.

Would it comfort you to learn now that the surgeon is going to do his first ever surgery on your loved one with no supervision? Probably not! As you start to drag your loved one off the operating table, would you be calmed down by the surgeon telling you that he came top in his (theoretical) surgery class; and would you be even more comforted to know that being top of the class was norm reference assessed? In other words the whole class may have actually 'failed' the tests but adjustments were made on the grounds that the test was 'too hard' and therefore the top students were given an 'A' grade, whereas had they been in the previous year, they would have failed the course? [to be even more concrete about this – the surgeon was top of his class but he only scored 23% on his paper].



I know I wouldn't be happy at all, and I would continue to take my child off the table and then search for a surgeon who had demonstrable practical competence and experience.

For vocational teaching/learning this is relatively easy.

- Can the student turn the '*Vorotex 3000*' lathe?
- Can the student nurse successfully extract blood with minimal blood loss or pain to a patient?
- Can a student successfully grow a cabbage in a hydroponic solution within a specified time period?

This does not however, mean that other disciplines cannot also be assessed using a criterion approach. It may require a bit more thinking as to what the actual criteria actually are. So for instance:

- ◉ A 10 year old should be able to recite successfully the whole mathematical times table without error by the end of the year (up to 10 times table).
- ◉ An 11 year old can answer a set of 100 multiplication questions that test the times table up to 10, out of sequence and in a time limit of 10 minutes with only 5 mistakes.

These are demonstrable, objective 'criteria' or 'benchmarks'.

Criterion Reference Assessments speak to a number of different advantages:

- ◉ They are not intelligence theory bound, either the child/student can do the task, or they cannot. If they can (or cannot) this says nothing about their presumed intelligence (or lack of it).
- ◉ Potentially all the class can achieve an 'A' grade (because they can all do the specified tasks) – equally well they could all get an 'E' grade (because they can't do the tasks).
- ◉ The criteria are task orientated with the intention of their being relevant to 'real world' tasks.
- ◉ The tasks are as objectively defined as possible, this means that an assessor can explain the assessment method to another assessor. No 'hidden' expertise is required as long as the assessment is adequately defined.
- ◉ The tasks are 'open' which means that they are open to scrutiny. A colleague, or indeed the community can query whether the specified criteria are relevant – or not.

There are of course, some negative aspects to this approach:

- ◉ Normally it takes more time and resources to construct the criteria – it's easier just to say "trust me I'm an expert and I know who is 'good and who is 'bad'!"
- ◉ It is sometimes intimidating to 'expose' your criteria, because of course they could be wrong, or inadequate, or improved upon.
- ◉ Not every task can be made into an easy to measure objective criteria, and some perhaps not at all.

Authentic Learning

Think for a moment what we actually want our schooling to give our children. I recently asked some of my children in class why we taught them maths. The answer from most children was 'because it is in the syllabus'. This surely is not the case. Surely we teach them maths because this is a genuine tool that they need to utilise in their everyday adult life? There is a whole educational movement based around the premise that we need to construct an educational programme around authentic tasks that are a better reflection of what children will encounter in the 'real world'[1].

[1] Check out these resources that are specific to authentic learning:
http://net.educause.edu/ir/library/pdf/ELI3009.pdf, although specific to distance learning, still good explanation.
http://net.educause.edu/ir/library/pdf/ELI3017.pdf, why students value authentic learning.
http://web.media.mit.edu/~mres/papers/authenticity/authenticity.pdf, paper talking about different 'authenticities.
http://www.authentictasks.uow.edu.au/, project done to promote authentic learning.

Validity

Validity is a technical term in psychometric assessment (the evaluation measures that psychologists use) that tells us to what extent a measurement 'really' measures the thing that we're interested in. In the previous section, I ended up by suggesting that IQ may not 'really' measure something that we are really interested in as employers of our future employees. In this regard IQ is not a great predictor (at best) of employee productivity or capability and is therefore said to have 'low validity'. IQ is however, a good indicator of future scholastic success in formal schooling, in which case IQ has 'high validity' when considering a child's future school performance. There are different types of validity, some of the boundaries are a bit 'fuzzy', which are summarised in the following table:

Type	explanation
face validity	I always title this 'superficial' validity in my head. It 'looks' like it's measuring the right thing. I want to test you on your maths ability, so I give you loads of maths problems.
internal validity	This has more to do with the validity of the test construction within itself, that is a randomised half of the test correlated very highly with the remaining half, regardless of how many times you randomise it.
construct validity	The degree to which a test measures the actual 'thing' that is of interest. Not always easy as we've seen with the IQ test trying to measure someone's 'intelligence', in part because we don't really have a solid definition of intelligence.
criterion validity	I believe that this is another variation of 'construct' validity, but strictly it's defined as the ability of the test to correlate highly with other acknowledged unrelated tests or measures that supposedly measure the same thing. I seem to remember learning this as 'convergent validity'.
ecological validity	This is the measure to which the test item correlates with 'real world' performance. A child may know about percentages, but can they translate that into compound interest rates on house mortgages?

The first two definitions are ones that are the easiest to show, or demonstrate. Internal validity is perhaps a little esoteric in that it has more to do with a statistical function, but given the right evaluation tools and enough sampled data, it's possible to tweak the results (take all the answers to questions 12, 18, 119, 154 & 155 – say), and see if the internal validity improves. Unfortunately they are also the ones that are of possibly the least importance.

The validities that we're most concerned about are really the last three: construct, criterion and ecological validity. These are the ones that speak most directly to the idea of authentic learning.

Authentic Assessment[2]

Validity has been given a newer title, one could think of it as a parallel invention from the more statistical sourced term 'validity' but it deals with the same issues. Specifically, does the assessment tool really measure how a person has understood their own learning – and ideally does it translate into real world activities.

In reality, these assessments are often more difficult to construct, require imagination to construct well as they often require complex task completion across a number of different disciplines, probably including ones that you as a teacher/lecturer/facilitator have not been specifically tasked to teach.

[2] There are some fabulous internet pages devoted to defining authentic assessment that I am doing here:
http://jonathan.mueller.faculty.noctrl.edu/toolbox/
http://www.eduplace.com/rdg/res/litass/auth.html
http://pareonline.net/getvn.asp?v=2&n=2

What is the Psychology Behind Criterion Referenced Assessment?

By now, you might be thinking 'hang on, this isn't psychology, this is an educational philosophy that he's trying to foist on us!'. Whilst this is partially true (to my mind criterion referenced assessment is so much more superior to norm referenced assessment), I think the reason for me explaining so much about this assessment approach has more to do with being a viable alternative to the norm referenced approach. The norm referenced approach has an apparent basis in psychology research, but I've tried to show that this psychological basis is flawed. Our psychological knowledge is so much more sophisticated since the 1920s and 1930s when these approaches were first being formulated. Educational practice however, has yet to catch up. In this regard criterion referenced assessment does acknowledge the progress in psychological theory.

- It acknowledges that the normal distribution of scores is circumspect in mental abilities. Since we still don't know (i) how to define intelligence properly, and therefore (ii) don't know how to measure it and therefore (iii) have no idea of it's actual distribution amongst our populations – best not to design an assessment methodology that relies on this information.
- It acknowledges the statistical distinction between relative position in a class, versus one's actual abilities against an absolute yard stick (not so much psychology, but really the statistics that is often used in psychology). Although many times the results maybe the same, for mission critical tasks (neurosurgery, atomic power plant operations, international diplomatic peace negotiators), we're best to rely on the 'absolute' task competency rather than the relative competencies.
- Finally there is a psychological theory that does supports this approach (but in a slightly sideways fashion) in that the '*Communities of Practice*' approach that we covered last week. This has been explored particularly by researchers Brown, Collins & Duguid (1989) who state that the learning occurs naturally along the lines of being placed in the situation in which the tasks are actually required. Brown et al calls this *situated cognition*.

Backwards Design

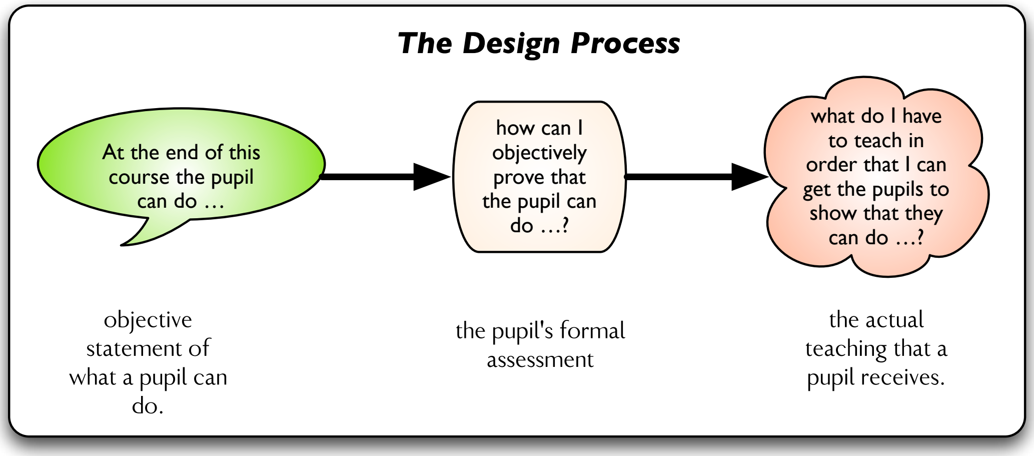
I wanted to end this lecture with an approach that I believe is not really related to psychology, but it does incorporate the criterion referenced/authentic learning/authentic assessment approach very well. So this is really 'practical' advice as to how to incorporate this in the 'real world'. This has to do with an educational design approach that I've labelled 'Back to front design' but apparently someone else beat me to the labelling and called it 'Backwards Design'. It's called this because you start with the end goal, and then work progressively backwards until you end up with your teaching tasks. This approach can be thought of in three phases.

(I) Defining what the 'end goal' is (which is the same as the criteria, in criterion referencing)

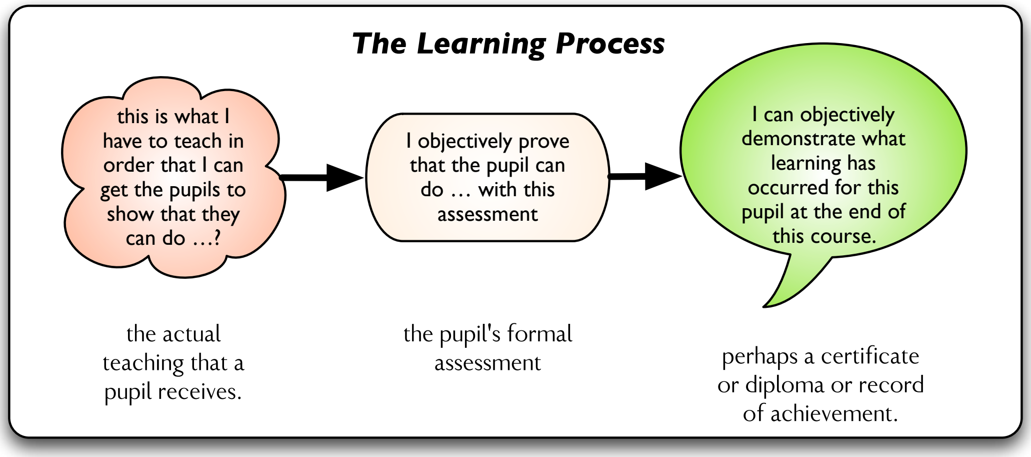
(II) Defining how you would prove that a student/pupil had achieved the end goal in an objectively measured way (this is really the assessment).

(III) Finally figuring out how you would teach your students/pupils towards this assessment.

This is show in the diagram below.


The Design Process
At the end of this course the pupil can do …
how can I objectively prove that the pupil can do …?
what do I have to teach in order that I can get the pupils to show that they can do …?
objective statement of what a pupil can do.
the pupil's formal assessment
the actual teaching that a pupil receives.

Of course, it is ‘back to front’, in that one starts the process in terms of thinking about the future and what happens at the ‘end’ of the line. Once the design has been completed you actually play the actual teaching and learning forward so now it looks like this.


The Learning Process
this is what I have to teach in order that I can get the pupils to show that they can do …?
I objectively prove that the pupil can do … with this assessment
I can objectively demonstrate what learning has occurred for this pupil at the end of this course.
the actual teaching that a pupil receives.
the pupil's formal assessment
perhaps a certificate or diploma or record of achievement.

The result is a logic that is a bit like looking at a map, finding out where you want to end up on the map, and then plotting your course from where you are presently. The backwards design process can be done at any level, from an individual lesson (slightly tedious), through to a complete year’s course (for instance I used this approach to design T5303), or even a complete curriculum (as we’ve done in the school that I also work at).

I’m sorry that there’s no empirical research that demonstrably shows that a backwards design process is ‘superior’ to a conventional method of designing a teaching programme. All I can tell you is that when you use this approach, criterion referenced assessment is the natural path to go.

Summary

There is some considerable psychology behind assessment in conventional educational practice. Unfortunately the psychological ‘wisdom’ is for the most part, the same underlying logic that was prevalent originally when psychology embraced statistics particularly around the 1930s and 1940s. The notion that all human trains fall on a ‘*normal distribution’* is considered a self evident ‘truth’. This provided the basis for forming a ‘*norm referenced assessment*’ system.

Since then, however, we have many reasons to distrust this self evident truth. At best, we should honestly state that we do not have enough discriminatory evidence to tell us whether this is a valid approach, or another is better, or perhaps some radical alternative way that makes all current approaches redundant.

So current psychological theory suggests that a reliance on the 'norm referenced assessment' is not currently appropriate.

Psychological theories do not tell us what an appropriate alternative could or should be. However, the most common alternative to norm referenced assessment is *criterion referenced assessment*. Criterion referenced assessment has the advantage of not being dependant on any psychological theory – rather psychology is neutral on this form of assessment.

References

Brown, J.S., Collins, A., Duguid, P (1989) Situated Cognition and the culture of learning. *Educational Researcher*, **18**, 32-41.

Shaffer, D. W, Resnick, M. (1999) "Thick" Authenticity: New media and authentic learning, *Journal of Interactive Learning Research*, **10**, 195-215.

Glossary

Norm referenced assessment	assessment that places people in a relative order from highest to lowest scoring. Cohorts of grade categories are based on a 'normal distribution curve'.
Criterion referenced assessment	assessment that establishes if someone can perform at an objectively stated standard of performance.
Normal distribution	a frequency distribution that has a 'bell shaped' curve when many points are plotted.
Standard deviation unit	the average distance away from the arithmetic mean in a distribution of population scores.
Frequency distribution graph	a graph that always has categories of the variable of interest on the horizontal (x-axis), and the absolute number or percentage of people that fall into those categories on the vertical (y) axis.
Skewed distribution	When a frequency distribution has a disproportionate number of people in a category that is not at (or close to) the median score. The graph normally looks asymmetric as if it's been pushed to one side or another.