oxfordenglishtesting.com

The Oxford Online Placement Test: The Meaning of OOPT Scores

Introduction

"The Oxford Online Placement Test is a tool designed to measure test takers' ability to function communicatively at different levels of English language proficiency according to the Common European Framework of Reference (CEFR)"

This paper explains how the OOPT *measures* ability, and how the results can be *interpreted* in terms of the CEFR. The questions were designed to ensure that each one tests a part of what it means to function communicatively in English (the principles behind this are described by James Purpura in another paper on this site *The Oxford Online Placement Test: What does it measure and how?*). For an effective assessment tool, however, we need to ensure that all of the questions work together, providing consistent information about the student's proficiency, and giving us an accurate and reliable result each time the system is used. There are four main steps necessary to guarantee this:

1 Constructing the OOPT Bank
2 Measuring ability
3 Interpreting scores
4 Assessing accuracy

1 Constructing the OOPT Bank

Initial analyses

Before test items can be added to a bank they must be checked statistically, and calibrated. Using traditional item analysis techniques, all of the potential OOPT were vetted to make sure they were not too easy or too hard and were working as expected.

Pretesting

Then, formal pretests were carried out in schools. Typically around 200 students from a dozen or more schools, and with a wide range of first languages, answered a test of round about 50 items, of which most were new but at least 10 were what are known as 'Anchor' items. The analysis used a Rasch latent trait model, which provides more sensitive checks on the behaviour of each item and a measure of its difficulty that is usually stable across different samples of students. Again, any items that didn't function as expected were removed, as was any student whose pattern of responses was odd (though there were very few of these). Attention then turned to the Anchor items.

These were items that had already been calibrated into the Bank and were now used to link the new items to it. First, we checked that they were stable, that nothing odd had

happened that might make one of them easier or harder for this group of students than it was when first calibrated. Diagram 1 below shows how this was done.

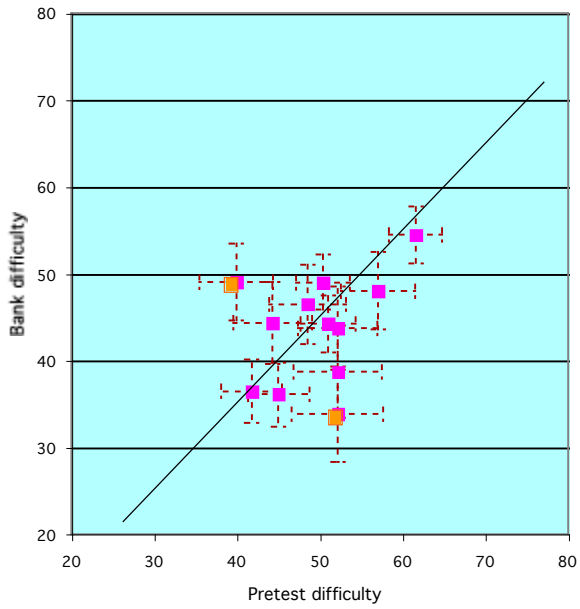


Diagram 1: Checking the Anchor items – 1

A point was plotted on the graph for each Anchor item showing how its new 'Pretest difficulty' compared to the 'Bank difficulty' from earlier analyses[1]. Ideally, the items should lie on a straight line, meaning that they had maintained the order of difficulty they showed before. In this case, two items (shown in orange) were clearly unstable; one seemed much more difficult than it had done before, and another one seemed much harder.

It is quite common for an item to behave unpredictably like this, and this is one reason why tests are as long as they are! The two items were removed from the analysis, and the graph re-plotted. The result is shown in Diagram 2.

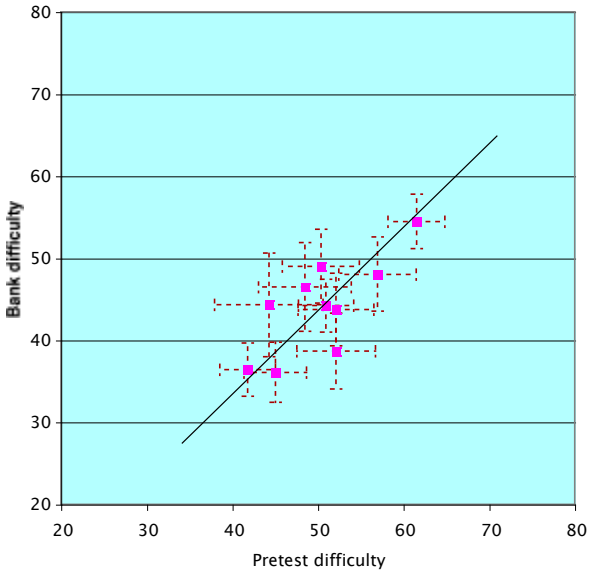


Diagram 2: Checking the Anchor items – 2

1 Note: the numbers on the axes of these graphs are internal numbers used by the OOPT system. How they are converted to the scale used externally for reporting results will be described later.

The dotted lines give an idea of how close the items need to be for the system to generate reliable results; in this case all of the items lie more or less within these limits showing that we now had good enough evidence to calibrate the new items accurately into the Bank.

Linking

When we analyse a pretest on its own we find out how difficult each item is *relative* to the other items in that analysis, but not how difficult they are relative to the Bank. To find this, we need a link between the new 'Pretest scale' and the established 'Bank scale': this is what the Anchor items provide. The procedure is illustrated in the next two diagrams.

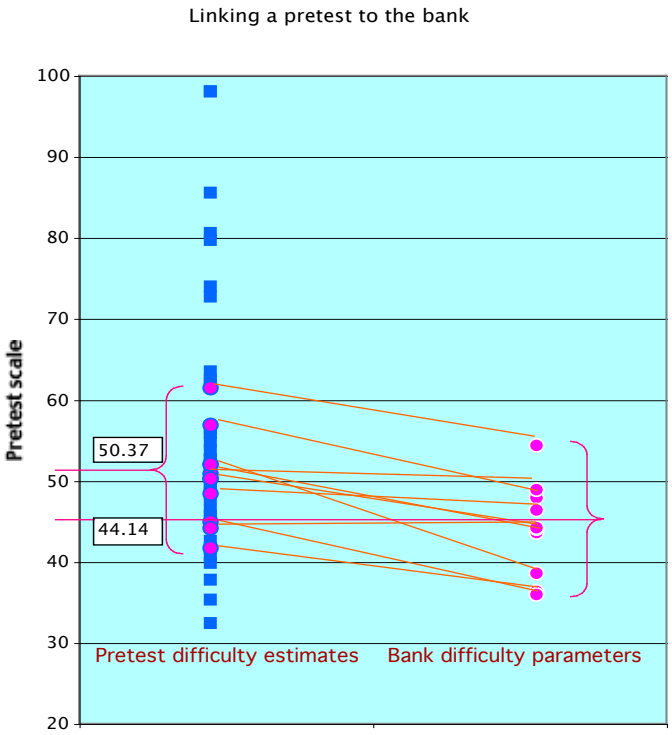


Diagram 3: Linking a pretest to the bank – 1

In Diagram 3 all of the items in the new pretest *including the Anchor items* are shown on the left, sorted into their order of difficulty, with the hardest at the top: these are labelled "Pretest difficulty estimates". The Anchor items are shown in pink. The 'Pretest scale' is a temporary scale for measuring difficulty, derived only from the pretest data. On this scale, the average difficulty of the Anchor items was found to be 50.37 units.

However, we already knew the difficulty of the Anchor items on the established 'Bank scale', and these are plotted on the right side of the diagram and labelled "Bank difficulty parameters". On that scale, their average difficulty was only 44.14 units. Since they are the same items, we can assume that, on average, they have the same difficulty. The difference between the two averages, 6.23 units, was the 'shift' required to bring the Pretest scale into line with the Bank scale.

We then deducted this shift from the 'Pretest difficulty estimate' for *every* item in the pretest – not just for the Anchor items. Diagram 4 shows the result of doing this. The lines linking the two values for each Anchor item were now *on average* horizontal: the link was complete.

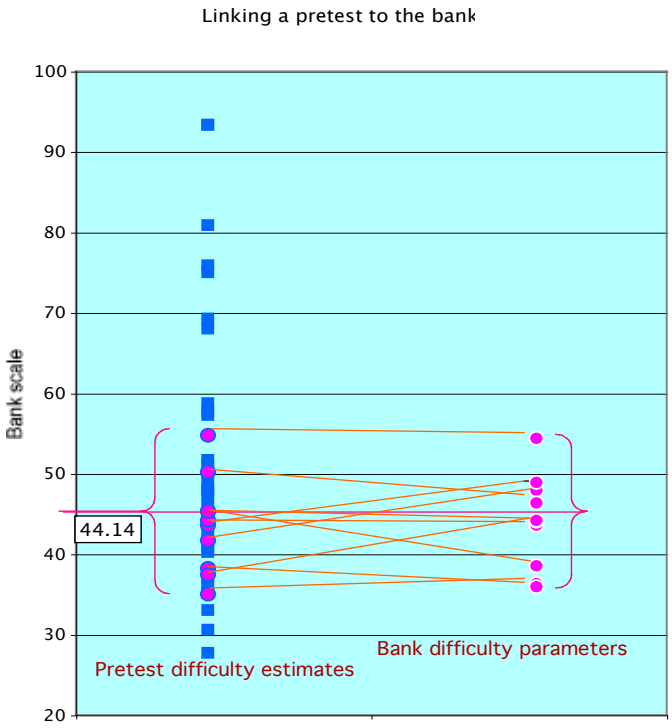


Diagram 4: Linking a pretest to the bank – 2

Linking relies on the average difficulty of the Anchor items, and it is important that this be stable. For the OOPT, items were only used as Anchors if they passed every earlier quality test, and Anchor items always made up at least a quarter of the whole pretest.

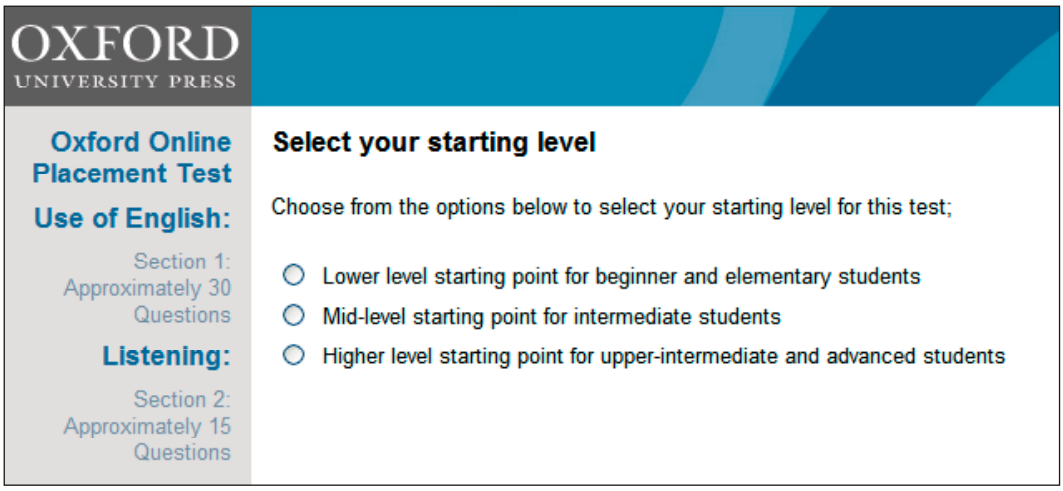
In this way, an initial Item Bank was built containing 462 Use of English items and 218 Listening items. This bank was used for the studies reported in the next two sections, to develop the OOPT system. At the same time, several hundred more items were developed, tested, and added to the Bank.

2 Measuring ability: adaptive tests in action

The Principle

Given a good coherent Bank of items, adaptive testing is an extremely efficient way of finding a student's score level; typically an adaptive test reaches the same accuracy as a traditional fixed test that is two or three times as long.

The OOPT gives the student the opportunity of selecting their starting level before they begin the test. The picture on the next page shows the point where the student may choose whether they wish to begin at a lower level, mid-level, or higher level starting point.



After the first item, which is chosen based on the starting point selected by the student, the OOPT chooses future items according to how successful a student is. Each time a question is answered correctly the system raises its estimate of the student's score a little, and so chooses a more difficult item to present next. This continues until the student gets one wrong, when the estimate and the selection are both lowered: very quickly, OOPT 'brackets' the student's performance level.

And because every item (after the first few) closely matches the student's level, the measurement is more accurate than any traditional test of similar length could be. No two students see the same series of items because each student's test is tailored precisely to their particular level.

Technical note

In fact, an adaptive system that followed this description exactly would quickly run into some technical problems well known in the academic literature: some items would be used far more often than others, certain strings of items would often occur – in fact we would expect a student who re-sits the test to get precisely the same test every time. To avoid them, a certain amount of variation has to be built into the process of selecting the next item. OOPT avoids the problems by defining a zone around the student's ability estimate, and choosing the next item randomly within this range. There is no noticeable loss of efficiency with this procedure.

Some Illustrations

Diagrams 5 and 6 show how the OOPT system operates; they are based on the records of real students taking the pilot version.

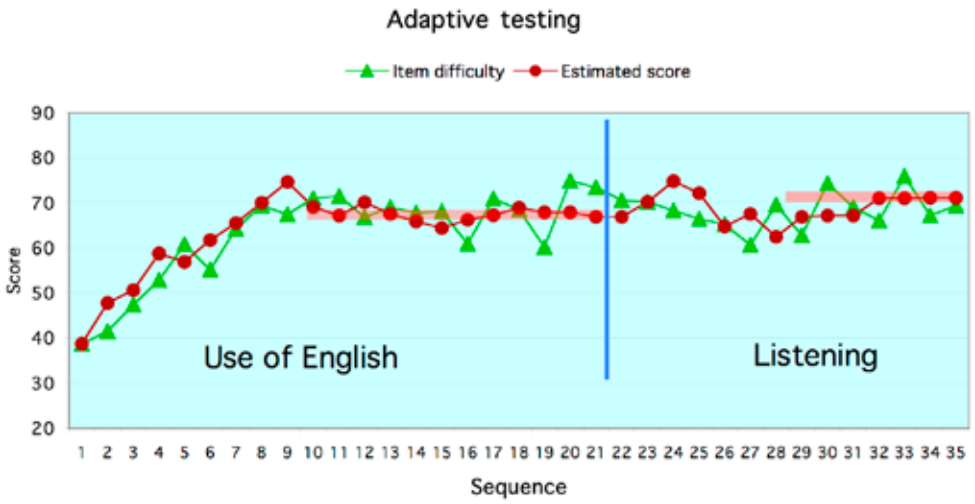


Diagram 5: Estimating ability and selecting items – *Student 1*

The system began by 'assuming' a low score, just below 40, for *Student 1* and setting her an item at that level. It doesn't really matter if this assumption is far from the truth – as it was in this case – for the system moves very quickly with the first few items. This student answered the first item correctly, and her score was revised upwards to about 50; an item a little above 40 was chosen and presented next. This progress continued as she got right all but one of the first eight items and her estimated score gradually rose to 75. But a couple of mistakes then proved that this was too high, and within a few items her score for Use of English levelled off just below 70.

With this established, the system went on to find her Listening score. For a starting point it used her Use of English score; after two successes, and two failures, the estimates quickly settled on a value just over 70 for Listening.

Notice that the system converged on a stable final score quite quickly, after only about 13 Use of English items, and after about 11 in the Listening section. This is what usually happens but a few more items are set, so that a test typically contains about 30 Use of English items and 15 Listening items; this ensures that the result for every student is accurate enough for the purpose the OOPT is designed to meet.[2]

2 Diagrams 5 and 6 use data from the pilot version of the OOPT, where the test contained about 21 Use of English items and 13 Listening items.

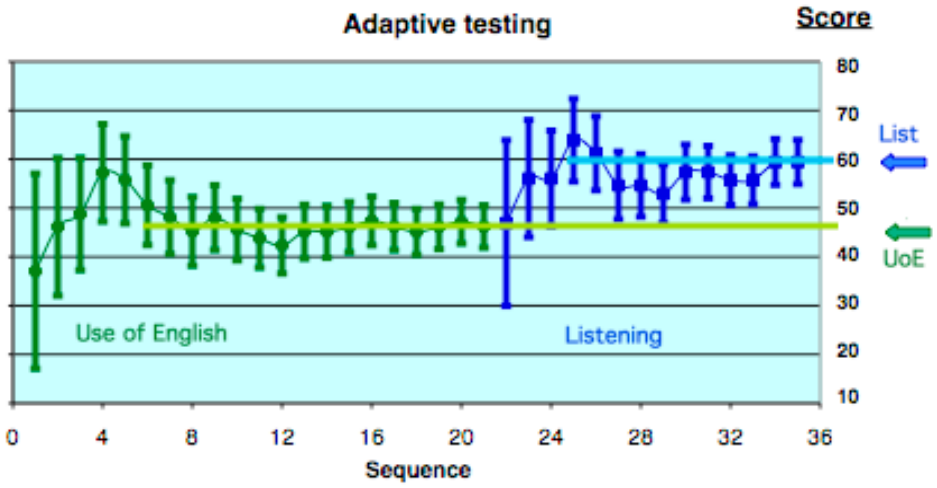


Diagram 6: Reducing uncertainty – *Student 2*

The pattern here was similar, as the initial score estimate was too low, and *Student 2* got the first three items right. His score for Use of English stabilised after about 12 items. In this diagram, the vertical lines indicate how much uncertainty there is about this score after each item has been answered.

At first, the uncertainty is quite large, at about plus or minus 20; after each additional item the uncertainty diminishes. There are usually 30 items in the Use of English, and after these the uncertainty has been reduced to less than plus or minus 5. Because students' Listening scores are usually quite similar to their Use of English scores, there is a little less uncertainty at the start of the second section, and the required precision is reached after only about 15 items.

The significance of differences

Note that *Student 2* seems better at Listening than at the grammar and knowledge of social language tested in the Use of English section. It's important not to assume that every difference between a student's two scores is significant, but this one is quite large and may have significant implications for his teachers or for his proficiency in different contexts. How large does a difference need to be to be considered meaningful?

In general, a difference of:

- 6 or less means it's not safe to think the student is really better at one aspect of proficiency than the other;
- 7 or more means it is likely that there is a real difference, though it may not make much difference in practice;
- 13 or more means there is a real difference that may be important for the student's future.

In any case, it is likely that other evidence about the student will be available, and the test results should always be considered alongside whatever else is known about them.

Conclusion

When approximately 45 items have been administered, the OOPT system is able to produce two scores for each student – one for Use of English, and one for Listening – with sufficient accuracy for student placement and similar purposes.

The next section will explain how these results are reported by the system.

3 Interpreting scores: link to CEFR

Equating

What do these 'scores' mean? What should a teacher think when Student 2 is reported to have scores of 47 for Use of English and 60 for Listening? The OOPT gives meaning to students' scores by converting them to the scale of the Common European Framework of Reference, as defined by the Council of Europe (www.coe.int/T/DG4/Linguistic/CADRE_EN.asp).

The conversion is based on the judgements of teachers from 30 schools in 11 different countries, who participated in the pilot phase of developing the OOPT. They were asked to rate their students on each of 50 general language use "Can Do" statements based on the Level descriptors in the CEFR. 638 completed reports were returned. With data like this it's important to check that all of the raters are interpreting the statements in similar ways, and a consistency analysis was applied, as a result of which some reports were rejected: 566 reports and 44 statements were used consistently and kept in the analysis. The number of statements ticked for each student indicated the Level the teacher considered the student to be at; most were rated in the range A1 to C1, although there were a few rated outside this range.

The method recommended by the Council of Europe for linking tests to the CEFR was not used. In their *Manual for relating Language Examinations to the CEF* they advise that a set of teachers should rate items, rather than students, asking, for each item: At what CEF Level can a test taker already answer the following item correctly? But this asks teachers to judge the difficulty of the items, and there is ample evidence in the research literature that even trained judges are not able to do this well – and especially when the items are multiple choice. In addition, of course, the OOPT is not a simple test with 40 or 50 items but a bank system with thousands of items, and it would not have been feasible to ask anyone to judge the difficulty level of all of them.

Instead, the scores of the students from the OOPT were matched to the teachers' estimates of the ability of the students so that the same number of students was assigned to each Level by the test and by the teachers. Linking the OOPT scale to the CEFR through students makes use of teachers' knowledge of their students and their standard. It is very common to use teachers' estimates like this to validate a test; here they were used to set standards for the OOPT through the method known as "equating percentiles".

The OOPT system reports each student's status on a continuous numerical scale:

Level	Score range
< A1	0
A1	1 – 20
A2	20 – 40
B1	40 – 60
B2	60 – 80
C1	80 – 100
C2	> 100

Thus, if a student is reported to have a score of 61 it means that they are judged to have reached Level B2, but not to have made much more progress towards Level C1. A score of 71 would indicate a student is comfortably in Level B2, and about halfway towards C1.

Aggregation

One issue has been glossed over so far. The OOPT generates two scores for each student, one for Use for English and the other for Listening. Each is derived independently, and reported separately. The system, however, is designed for *placement*, and it is generally not possible to place a student into two different classes at the same time. A single score for English is needed.

The answer is simply to average the two scores. Hidden in this is a decision, reached after considerable discussion, that there is no reason to do other than give equal weight to the two components. In future, there may be more than these two components in the OOPT: if so, then this decision will be reconsidered.

Because the two scores are reported as well as the aggregate, users may of course give different weight to the components if they wish. In some contexts, such as when English is being taught with an emphasis more on conversation than on reading and writing, it may make more sense to give extra weight to the Listening score. It should be borne in mind, however, that the scores are quite close for most students and that changing the weights will not make a great deal of difference.

4 Assessing accuracy

One of the diagrams used earlier showed how the "uncertainty" around a score reduces as more items are answered. It's important to understand this concept in order to avoid the risk of misinterpreting test results.

Measurement can never be perfect, and when we try to measure something both unobservable and as elusive as language proficiency there will always be a significant amount of uncertainty associated with the scores obtained. Some of this uncertainty comes from variability in the students: their alertness, concentration, or motivation can fluctuate from day to day, causing similar fluctuations in their scores, and there is not much that any test can do to overcome this effect. The other main source of uncertainty

with test results comes from a particular set of items that a particular student meets; would they have done better (or worse) if they had met different items? The OOPT uses an adaptive testing methodology that is much better at reducing this problem than any fixed test can be.

Before a student answers the first question, the system has no information about her and so is completely uncertain about her level. If she gets the first item right then it knows that her score is *probably* higher than the item's, but it doesn't know how much higher. To fix her position the system needs to find some items that she can't answer correctly *and* some she can – then her score should be somewhere between those limits. Each item contributes a little more information, but the ones that are much too difficult or easy don't tell us much: they are too hot or too cold for her. The best information for fixing her score comes from the ones that are 'just right' for her.

Anyone familiar with English children's stories might like to think of this as the 'Goldilocks strategy', after Robert Southey's tale of *Goldilocks and the Three Bears*, but the principle is fundamentally one of statistical theory. Adaptive testing constantly seeks items that are 'just right', according to the latest information about a student's level, and that therefore reduce the uncertainty as much as possible.

At the end of the process the remaining uncertainty is reported, conventionally, as a 'standard error'. This is a rather unfortunate term, since it seems to imply that mistakes have been made in measuring the student; since the purpose of a test is to give us a clearer idea of a student's current level of ability, a term such as 'residual uncertainty' would be better. At the end of an OOPT session, the residual uncertainty will normally be 5 units.

What then does it mean if a student is reported to have scored 65, with an uncertainty of 5? First, it means that the best estimate we can make is that his current ability is 65, well within B2, and about one quarter of the way towards C1. Second, we can be more precise about how uncertain we are, like this:

Score = 65 ± 5 Probability that the student's real ability is currently:				
below 55	below 60	between 60 and 70	above 70	above 75
2.5%	16%	68%	16%	2.5%

Despite the inevitable uncertainty that surrounds any measurement, we can be fairly sure this student really is able to operate at Level B2: there is only about a one-in-six chance that he has been wrongly misclassified and really is still in B1. But remember – if six students were all reported to be one standard error above a boundary, then it's very probably true that one of them really should be below it!

What should a teacher do if a student is reported to be very close to a boundary? The technical answer is to say that this student needs to be measured more carefully to avoid the risk of misclassification, and so the student should take another test. The standard error for the two tests combined would be reduced by a factor of √2, from 5.1 to 3.6, allowing a decision to be made with less uncertainty. However, the practical answer might be to make a decision anyway, but to keep that student under particular monitoring, or to consider other information that may be available, to help make sure the best decision is made.

A test result is the score a student gets on one particular occasion; it might be different on another occasion, and different kinds of assessment can also lead to different results. As mentioned earlier, test results should always be considered as one, relatively objective, component in the whole picture of the student's progress so far.

Traditional test concepts

Score

In a traditional test, a student's *score* refers to the number of items they got right: it is a measure of *how many* items they succeeded on. Adaptive tests work differently: the score is a measure of *how difficult* were the hardest items the student could succeed on. Technically, the score is equal to the difficulty of items the student has a 50% chance of getting right: it's a judgement rather than a count of anything.

One useful consequence, alluded to earlier, is that the "just right" strategy ensures that every student gets about half of their test items right and half wrong. There will be little or no impact on their motivation or confidence during the test, and so one source of unreliability in the scores will be removed.

Reliability

The traditional concept of *reliability* is not appropriate for adaptive test systems, since each student actually takes a different test from everyone else. This means that there is no way to calculate the favourite indices of reliability, Cronbach's alpha or KR20. These statistics are, however, often misused since they depend heavily on the range of ability in the sample of students used to measure them; it is easy, and not uncommon, for test constructors to 'cheat' by including some students with inappropriately high or low ability so as to inflate the statistic.

However, most of the useful work done by the traditional concept of reliability is done by the idea of *standard error* or *residual uncertainty*. As described above, this addresses the critical question of accuracy: *How sure can I be that this student's real current ability has been measured?*

There is another useful notion often described as *test-retest reliability*, or *equivalent forms reliability*, or *the coefficient of equivalence and stability*. Whichever form of it is used, the idea is that a group of students take the test and then take either it or an 'equivalent' form again after a certain time. This coefficient then shows how stable the students' scores are on the two different occasions and, if equivalent forms were used, across different sets of items. The result will reflect any differences between the forms, and any changes (such as learning) that occur in the time between the two administrations.

Because the result is a correlation coefficient, it is again open to abuse because it is strongly affected by the range of ability in the students. Nevertheless, it addresses an important question: *How sure can I be that the student would get the same score on a different occasion?*

With an adaptive system, any student re-taking a test will, of course, receive an 'equivalent' form, rather than seeing the same test twice. OUP is currently collecting

data to evaluate the OOPT in respect of its coefficient of equivalence and stability. Early indications are good: Diagram 7 shows the first results, from a sample of 128 students in 5 schools.

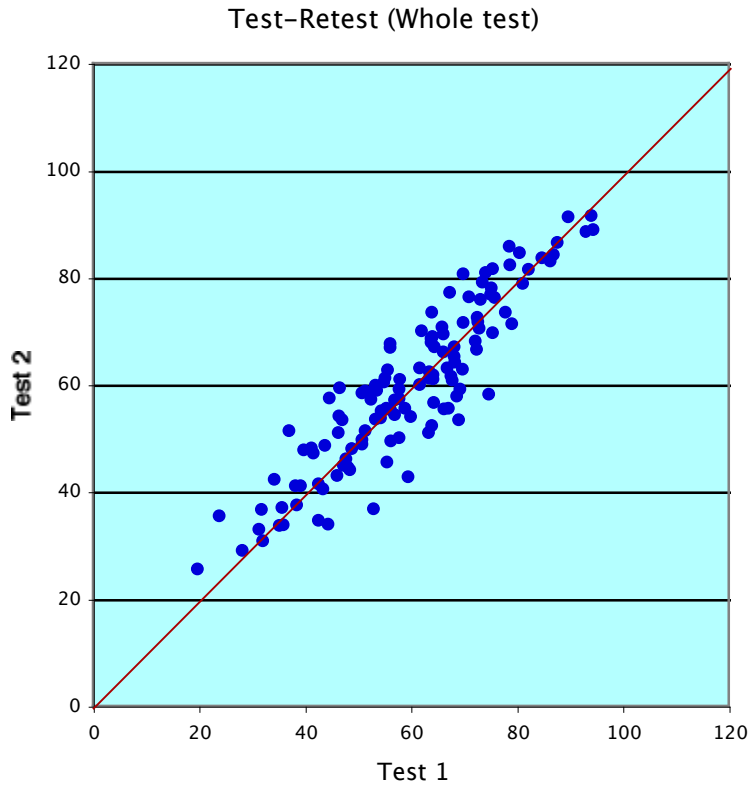


Diagram 7: Preliminary test-retest results

The correlations for these students were:

	Correlation coefficient	Standard deviation
Total test	0.91	15.30
Use of English	0.87	15.34
Listening	0.82	16.96

The diagram shows that these students ranged from about 20 to about 100 on the scale, and the standard deviations (which measure how spread out the students were) were between 15 and 17. In the full population used for setting the OOPT scale, scores ranged more widely – from 0 to about 120 with a standard deviation of about 22. It was noted earlier that correlation coefficients are affected by range; given that these students were less spread out than the whole population, we can be reasonably confident that the OOPT is highly reliable from the test-retest point of view.

Validity

There are two aspects to the traditional concept of validity – one intrinsic and the other extrinsic.

Measuring what we say we are measuring

This notion of validity is the one captured by the classical definition that a test is valid if it measures what it claims to measure.[3] To meet this requirement, validity has to be designed into the test by writing items in such a way that we don't just know how difficult they are but we also *understand* their difficulty. The key to this intrinsic notion of validity is that the student's current level of language proficiency is what *causes* their score to be high or low. Validity is a theoretical property of the test, achieved by the deliberate and careful writing of the items, and by a monitoring strategy that deletes items that turn out to be much more, or less, difficult than expected, since they show where the writers have failed to control the difficulty of the item.

How the OOPT has sought to ensure intrinsic validity is described in detail by James Purpura in the paper already referred to, *The Oxford Online Placement Test: What does it measure and how?*

Use and misuse

The extrinsic aspect of validity relates to the validity of the use to which the test results are put. No matter how well the test constructors have carried out their part of the process – how valid the test is – they cannot prevent other people from misusing it. If this happens, then invalid use is being made of a valid test.

The OOPT has been designed, in the theory that underlies its items and in every aspect of the measuring system, to serve one primary purpose – the accurate placement of students learning English into a class or a course of teaching that is appropriate to their current needs. Anyone using it for this purpose can be confident that it is as valid for that use as possible.

Of course, it could be used for other purposes, some only slightly different such as to monitor progress through a course of learning, or others that are very different, perhaps as a selection test for employment. The responsibility for maintaining validity then lies on the user, who must be prepared to defend the view that this test, designed on these principles, serves that purpose well enough. Of course, the test is delivered using the Oxford English Testing Learning Management System and this means that any institution can collect additional evidence to support their specific interpretation of scores.

This is the essence of extrinsic validity: is the test right, or right enough, for the purpose I want to use it for? Rather than *validity*, some people prefer to call this concept *utility*.

Note about the author

Alastair Pollitt is a member of the UK OFQUAL Advisory Committee, a visiting fellow at Cambridge University's Research Centre for English and Applied Linguistics and is internationally recognised in the field of educational measurement.

3 In the UK, a successful series of advertisements for paints and preservatives has led to this form of validity being known as: 'the test does exactly what it says on the tin'.