Statistical tools for language assessment

Most of the day to day assessment work carried out by teachers in the classroom requires little statistical knowledge. Sophisticated analysis tools that work well when used with hundreds or thousands of test takers are not generally of much real help to a teacher who works with classes of twenty, thirty or sixty students. Language teachers can become very effective as assessors without becoming statisticians.

On the other hand, language teachers do benefit from at least a basic awareness of statistics – of the principles that inform them, if not the details of the calculations. This knowledge can help them to understand the meaning of external test scores as well as helping them to improve the quality of the assessment materials they use and to carry out effective classroom research.

Naturally, statistical tools and techniques underpin the controlled world of standardised testing. Teachers who take on responsibility for preparing tests and assessment systems at an institutional level, researchers who make use of tests and, above all, professionals who wish to specialise in this field will need to develop a level of statistical knowledge that can be quite challenging for people who have chosen a career in languages over mathematics or science.

This guide is only intended as a preliminary introduction to some of the tools available and their functions in operating a language assessment. Doing justice to the use of statistics in language assessment even at a basic level requires a book in itself. Happily there are some excellent entry level texts available complete with practical step by step exercises. Recommendations on further reading are made below and in Exploring Language Assessment and Testing. For people intending to develop and evaluate their own tests or assessment systems, a book on language testing statistics will make essential preliminary reading.

Interpreting assessment scores
Statistics can provide helpful ways of describing and summarising people’s performance on one or more assessments.

How well has this assessee or group of assessees performed?

How many questions has the assessee/ group responded to correctly?

Has this assessee/ group performed better or worse than other assessees?

Has this assessee performed well enough to move on to the next unit?

We are all very familiar with the idea that a student who gets ‘full marks’ on a test has performed well. 50 questions, 50 correct: good result.

It may be less clear how well a student has performed if they score 35 out of 50. Is that a good score or not?
To answer that question – to interpret the score - we need to think carefully about the purpose of the assessment and the expectations we have of assesseees.

In proficiency assessment, we might use a standard setting procedure (see Exploring Language Assessment and Testing, Chapters 5 and 6) to decide how many points on the test the assessee would need to score in order for us to recommend that they could be employed as a tour guide (if we were also convinced that the content of the test provided sufficient and relevant evidence for this purpose).

In educational settings, we might be more interested in whether the assessee was performing at the level we would expect for a student in her year group with a similar number of hours of language study. 35 out of 50 might be an excellent score for a beginner learner on a difficult test or a very poor score for an advanced learner on an easy test. Even 50 out of 50 may not appear as outstanding if all the students in the class scored 45 or higher.

**Norm referenced score interpretations**

Deciding what is easy or difficult often comes down to what is typical for learners of a certain age and language learning experience. Parents become anxious if their child is not performing as well as others in his class. Teachers worry if their class is generally performing below others in the school or below the national average. Ministers of education are horrified if international test results indicate that students in their country perform much worse than those in neighbouring countries.

In all these cases, expectations about how well learners should perform are shaped by what is typical or normal for a certain population. This is ‘norm referenced assessment’ (see the section on relative and absolute score interpretations in Exploring Language Assessment and Testing, Chapter 4).

**The normal distribution**

A key statistical concept is used when interpreting test scores in these ways: the normal distribution.

The normal distribution is something that is very widely observed in nature. If you look around you in any busy street you will notice a few very tall people and a few who are quite short, but most will be quite close to average height. Similarly, if you give a language test to people who have all been studying for a similar length of time, their scores will tend to cluster around the average or *mean* score for the group. We can find out about this distribution of results by counting how many people fall within certain ranges of height (how many are 141cm – 150cm tall, how many 151cm to 160cm and so on) or how many people obtained each of the possible scores on a test. In fact, if you make a lot of observations and plot a graph showing the distribution of results, the outcome is very likely to look like this (Figure 1):
In the case of Figure 1, we are looking at the results of a test consisting of 200 items that was administered to 2,000 people. Each bar represents the number of people obtaining scores within a range of 20 points. Almost 700 people scored between 81 and 100 points. Slightly fewer – around 650 – scored between 101 and 120. Nearly 300 people scored 61 to 80 and a similar number scored between 121 and 140. Very few people scored less than 40 or more than 160.

If you draw a line connecting the top of all these bars, you will get this (Figure 2):

Figure 2
This shape — the famous ‘bell curve’ — is symptomatic of a normal distribution of scores. In reality, not many sets of results will come out looking quite as neat and bell-like as this — especially when numbers of assesses are quite small. The curve will probably be less even and the sides of the bell may be steeper (the scores are tightly bunched around the central point: the mean score) or flatter (the scores are more evenly spread out). Sometimes there will be a rather steeper slope on one side and a gentler slope on the other: the peak is a little ‘skewed’ away from the central point on the curve. However, as long as these distortions are not too great, the distribution of scores can be treated as normal.

The great value of a normal distribution for interpreting test results is that it has certain known properties. If we know just two facts about a set of scores that is normally distributed, we can use the attributes of the normal distribution to infer a good deal of potentially useful information.

These two facts are the mean and the standard deviation.

Means and standard deviations
Work through Task 1 below to see how the mean is calculated from a set of test scores.
## Task 1

### Ranks

Ten people took a 100 item language test: the *Test of English Achievement* (TEA). These were their scores.

Amy 53, Ben 77, Carla 74, David 25, Emma 30, Freddy 34, Gladys 13, Harry 23, Irma 90, James 55

Who was the top scorer?
Who was the lowest scorer?

### Means

The mean (or average) is calculated by adding together all the scores on the assessment and dividing the total by the number of assesseees.

**Step 1** Add up all the scores. What is the total of all the points for all the assesseees?

**Step 2** Count the number of people. How many are there?

**Step 3** Divide the answer to Step1 by the answer to Step2.

What is the mean?

You can check your answers at the end of this guide.

The mean provides a useful picture of how well a group of assesseees has performed. If the mean is very high, the material was relatively easy for them. If it is low, the assesseees generally found the material difficult. This can have implications for the assessment (maybe it was a bit too demanding at this stage?) or for the assesseees (maybe they need to spend longer studying this aspect of the course?).

Although the mean usually gives us a helpful sense of what is typical for the assesseees and how well the group performed, it can be misleading. If three people took a 20 item test and all scored 10 points, the mean would be 10. In this case, the mean would be a good reflection of how well all of the students performed. Alternatively, if one person scored 16, another 14 and the third 0, the mean would also be 10, even though nobody actually obtained 10 points. The mean wouldn’t reflect how well any of the students performed. To gain a fuller picture, we need to know more: We need to know about the *dispersion* of the scores or how far they are spread out around the mean. The standard deviation gives us an indication of this.

To calculate the standard deviation, we need to know how far away each score is – how far it deviates – from the mean. In the example of the TEA test in Task 1 above, we found that the mean was 47.4. Irma scored 90. Her score of 90 is 42.6 points away from the mean (90 – 47.4 = 42.6). Gladys scored 13. Her score is 34.4 points away from the mean (47.4 – 13 = 34.4).

Calculating the standard deviation is more complicated than calculating the mean (it is clearly explained by Bachman 2004 among others), but it provides an overall indication – something like an average distance – of how far students’ scores tend to fall from the mean. In this case, the standard deviation is 25. If all ten students had scored 50 points, the standard deviation would be 0: the mean
would be 50 and all the scores would be exactly equal to the mean – a distance of 0. If half had scored 0 on the test and half 100, the mean would still be 50, but the standard deviation would also be 50. If the standard deviation is a relatively small number, the scores are mostly very close to the mean: if it is a large number, the scores are more widely spread out.

Now for the magic of the normal distribution. Having calculated the mean and standard deviation, we can get immediate insights into just how well any individual assessee has performed relative to all the others.

![Figure 3](image)

We know that when scores are normally distributed, just over two thirds of the population of assesses (68.2%) score between one standard deviation above or below the mean (see Figure 3). But just 0.01% of the population is likely to score more than three standard deviations above or below the mean.

In some circumstances, it may be more useful to know a student’s position in relation to the population as a whole than to know how many points he or she scored. For example, we might want to know if a boy is performing better in languages or in science. Of course, we can’t directly compare the two as they are very different subjects. However, we can see how well the boy is performing relative to a reference group of learners at the same stage in their education.

If the boy obtains a score equal to the mean for science, he has performed in this subject as well or better than 50% of the students in his year group. This is shown by the cumulative distribution or blue arrows above the horizontal axis. If he also scored two standard deviations above the mean for languages (i.e. as well or better than 97.7% of the year group), he is probably more gifted as a linguist than as a scientist and he might be advised to consider a career as a language teacher. If his score for languages is two standard deviations below the mean (as well or better than 2.2% of his year group), intervention – such as extra classes - may be needed to bring his performance up to the level expected.

**Standardised scores**

Sometimes, knowing how far the score is above or below the mean is considered to be more useful than knowing the raw score (the number of points or marks out of the total available). Raw scores can be transformed into *standardised scores*. These can take different forms. Standardised scores
that use standard deviations as a unit are known as z-scores. The mean, whatever it may actually be, is set at 0. A score of 1 is one standard deviation above the mean: a score of -2 is two standard deviations below the mean.

For example, a test had a mean of 34.2 and a standard deviation of 4. Kim scored 25 points. To convert this to a z-score, the mean becomes 0. The score of 25 is 9.2 points below the mean or 2.3 times the standard deviation (2.3 x 4 = 9.2). Kim’s z-score is therefore -2.3.

Unfortunately, being told that you scored -2.3 on the test can be confusing as well as depressing, so z-scores are not often reported to assesses. Instead, T-scores are used. Here the mean and standard deviation are set to some arbitrary figure such as a mean of 50 and standard deviation of 10. The score of -2.3 would become 27: 2.3 standard deviations multiplied by 10 gives 23. If 23 is subtracted from the mean (now set as 50), the result is 27. This now looks more like the kind of score students are familiar with.

Figure 3 shows the results for 2,000 test takers on the English Achievement Test – EAT. The mean on this test was 94.2 and the standard deviation was 23.8 (shown in the row marked ‘raw scores’ below the chart). The score on the EAT test of 118 is one standard deviation above the mean (94.2 + 23.8 = 118). This could be reported as a z-score of 1.0 or as a T-score of 60\(^1\). Referring to the cumulative distribution of test scores (shown by the blue arrows at the base of the chart in Figure 3) we can see that a score one standard deviation above the mean is in the 84\(^{\text{th}}\) percentile – or better than 84% of the test taking population. In fact, this student was Carla, who scored 74 out of 100 on the TEA test.

If we were to convert Carla’s scores on the EAT and the TEA to percentages, she would have 74% on the 100 item TEA test and 59% (118 out of 200) on the EAT. On the face of it, her performance on the TEA test would seem to be substantially better - by 15 percentage points. But is 74% on the TEA test really better than 59% on the EAT test? Or is it just that the EAT is more difficult? Perhaps 59% on the EAT actually represents a better performance than 74% on the TEA.

Provided that the means and standard deviations were calculated on the same population of test takers in both cases we could use the performance of the whole group to compare Carla’s scores on the two tests. In other words i) all 2,000 students took the TEA test as well as the EAT test and ii) we use means and standard deviations based on the scores from these 2,000 students when we compare Carla’s performances.

In fact, the means and standard deviations used in calculating standardised test scores are usually worked out on a reference population: a group who took the test when it first came into use. Scores on subsequent administrations are all reported in relation to those original results. The group taking the test today may all achieve scores that are higher than that original mean score: they may all achieve ‘above average’ results.

\(^1\) The mean is set to 50, the standard deviation is set at 10. One standard deviation above he mean becomes 50 + 10 = 60
Task 2
Remember that on the TEA test, the mean was 47.4 and the standard deviation was 26.4. Try converting Carla’s score of 74 to a z-score. What does this tell you about Carla’s performance on TEA and EAT?²

Using standardised scores in grading
A problem that often comes up for teachers is that scores from different assessments have to be combined to generate an overall result for a student over a term, semester or year’s work. Standardising scores offers one way of putting them all onto a comparable scale so that they can aggregated in a relatively objective way.

Let’s say we wanted to base an overall score for this semester on both the TEA and the EAT tests. If we just added together the scores from the TEA (out of 100) and the EAT (out of 200). The EAT would carry more weight in the final score than the TEA³. Standardising the scores equalises the contribution of each. On assessments with a relatively narrow spread of scores (such as tests of speaking skills scored on a rating scale) an excellent performance might be only a few points above an average performance while a good grammar test score could be 20 points above the mean. Standardising the scores gives more credit to the excellent speaking score.

Disadvantages of standardised scores
The downside of reporting standardised scores is that they are even less informative for users than raw scores about how well the assesseee performed the test tasks. They do reveal that Carla performed better than most of the learners in the reference group, but it is possible that all the test takers performed very poorly. Standardised scores alone do not reveal much about the knowledge, skills or abilities Carla actually demonstrated in arriving at her result. They do not tell us what an assesseee scoring 60 on the test can do that a student scoring 50 cannot.

These days, test providers often report Can Do statements alongside standardised scores. These statements may be based on student self-report data (most students scoring 60 report that they can understand newspaper articles in English) or on expert judgement about the functional abilities of learners at different score levels (experts consider that a student scoring 60 would typically have sufficient ability to follow academic lectures given in English). Users should be aware that tests reporting scores in relation to Can Do statements may not actually require test takers to perform the kinds of activity described.

Criterion referenced score interpretations
Another reservation that has been expressed about standardised scores is that they encourage assessment developers not only to exploit the properties of the normal distribution to facilitate interpretation, but to try to design assessments in such a way that they produce normally distributed

---
² Carla’s score on the TEA test is 74. The mean is 47.4. This is a distance of 26.6 points. (74 – 47.4 = 26.6) above the mean. The standard deviation of the test was 26.4 so Carla’s z-score was 1.01 (a T-score of 60): it is effectively indistinguishable from her performance on the EAT test.
³ Add the two scores together and divide by three to get an overall percentage: 118 + 74 = 192/300 = 64%. This is twice as close to the 59% on EAT than the 74% that Carla scored on TEA. The EAT has a bigger impact on her final score.
results. Instead of trying to find out whether or not students have mastered the intended content, their main aim is to spread students out and distinguish between them.

Spreading out learners is important for some purposes (such as placement testing – sorting students into different instructional groups). But education, it can be argued, is partly a struggle against the normal distribution. Unlike a proficiency test, education should not be about picking out the best and rejecting the rest, but of helping everyone to develop their abilities.

According to the concept of mastery learning covered in Chapter 7 of Exploring Language Assessment and Testing, detailed descriptions of what learners Can Do with the language should be the starting point for the design of useful, authentic educational assessments which provide clear evidence of whether or not the learner is really able to ‘understand newspaper articles in English’ or ‘follow an academic lecture’ or to do whatever else is covered in the curriculum. If at the end of the year every student shows that he or she is now able to do these things and so succeeds on a final assessment, this should be understood as a positive outcome. The assessment should not be rejected on the grounds that it fails to discriminate between higher and lower ability learners.

The distribution of scores on an instructionally sensitive assessment would not be normally distributed, but might be expected to look more like this (Figure 4):

![Figure 4](image)

Here a test – the Mastery of Unit Six Test (MUST) – was administered to a class of 50 students and scores were awarded out of 30 points. A pre-test based on the content of the unit that was about to be taught (Unit Six of the textbook) was given before the class started work. On this pre-test, most learners scored only a few points because they had not yet learned the material (if they had all aced the test, the teacher might have decided to skip Unit Six and teach something more challenging instead).

At the end of the unit, a post-test was given. This was another version of MUST that contained different tasks, but was equally difficult and was based on the same specifications as the pre-test. In other words, it also covered the points that were taught. On this test, most of the learners were successful and scored 20 or more out of 30. Unlike the information from the TEA and EAT tests, the results from this test could inform immediate classroom decisions. Most of the class performed well, but Bolormaa, who scored 17 out of 30, clearly needed some extra help with this unit. The teacher
looked at which items Bolormaa answered incorrectly and looked for ways of improving her understanding of the points involved.

The purpose of the MUST was not to help the teacher to spread the students out – to sort the ‘A’ students from the ‘F’ students – but to help the teacher to focus intervention where it is most needed. In this kind of assessment system, it should be possible for every learner to achieve all of the targets set and to achieve an ‘A’ grade for the course.

In practice, in language education it is very difficult to create instructionally sensitive assessments. Beyond the earliest beginner stages of learning, it would be unusual to find sounds, vocabulary, grammatical patterns, texts or other features that learners are completely unaware of before instruction. Because language is so complex and the different aspects of language ability that we can describe are so interdependent, mastery of linguistic features is rarely a straightforward matter. Learning gains tend to be gradual, variable and much less dramatic or clear cut over the short term than shown in Figure 4 above. In fact as we begin to experiment with using new linguistic forms, our accuracy in using them may actually get worse. However, the principle of instructionally sensitive assessment that helps to promote success in learning (rather than serving to identify winners and losers) is very attractive to teachers.

**Test reliability**

If we think about Carla, who scored 74% on the TEA test and 59% on the EAT test, we can represent her (percentage) scores on the two tests visually like this (Figure 5):

![Figure 5](image)

Alternatively, we can compare her performance on the two tests by representing the TEA test on a horizontal axis and the EAT test on a vertical axis like this (Figure 6):

![Figure 6](image)
Carla’s score of 79% on the TEA test is places her three quarters of the way along the horizontal axis, her score of 54% on the EAT test puts her just above half way up the vertical axis.

To compare Carla’s performance on the two tests with those of another student – Luigi – we can add his scores to the picture. He scored 35 on the TEA test and 17 on the EAT test. This shows quite clearly that Carla has done better than Luigi on both tests.

![Figure 7](image)

If both tests are supposed to measure the same thing – in this case it happens to be the range of language skills covered by the Year 10 school curriculum – we would expect high-flying students like Carla to perform well on both tests. Equally, we would expect Luigi, whose teachers report that he has struggled with the classes, not to perform as well as Carla. His low scores on both tests are consistent with each other and with what his teachers have said.

Now let’s add a third student to the picture. This is Kanya. She scored 25% on the TEA test and 77% on the EAT test. These results are not at all what we would expect based on the results from Carla and Luigi. Her score on the TEA test suggests that she is performing better than Carla – she is one of the top performing students in her year group. On the other hand, her score on the EAT test is lower than Luigi’s and suggests that she needs urgent help with her English language skills. Which conclusion is right?

![Figure 8](image)

The statistics alone do not provide the answer. It may be that we can explain the results using what we know about Kanya. Perhaps she is really a very high performing student, but was unwell when she took the EAT test and didn’t perform as well as usual. Perhaps she is really not a very able student, but was able to cheat on the TEA test by copying her answers from the very high performing student sitting next to her.

On the other hand, it may be that we can explain the difference in the scores because of characteristics of the tests. Perhaps there is something about the way the tests are made or about how they are scored that has led to these inconsistent results. Have construct irrelevant factors affected the scores? We may discover that both tests are tests of English, but the questions in the
EAT test are presented in Italian. Carla and Luigi both speak Italian, but Kanya does not: the TEA test is a much better reflection of Kanya’s English language skills than the EAT test.

Whatever the problems may be, the point is that inconsistencies in test results are like alarm bells. They show us that something unexpected has happened and indicate that we need to carry out an investigation to discover why.

---

**Ideal positive relationship**

![Correlation = 1.0](image)

**No relationship**

![Correlation = 0.0](image)

**Ideal negative relationship**

![Correlation = -1.0](image)

---

**Figure 9**

When two tests are both intended to measure the same construct (as in the case of the TEA and EAT), the relationship between them should look like the ‘ideal positive relationship’ in Figure 9. Students obtaining high scores on the TEA test would get high scores on the EAT test, those with low scores on the EAT would have low scores on the TEA. If we know the student’s score on one test, we could predict precisely what the score would be on the other.

In the ‘ideal negative relationship’ higher scores on the TEA test would be associated with lower scores on the EAT test. Of course, if both tests are supposed to measure the same construct (as with TEA and EAT), this should not be possible. But there are cases where a negative relationship is expected. For example, the first test is a language test, but the second is a questionnaire measuring anxiety about taking tests. Students who feel more anxious (high scores on the questionnaire) may get low scores on the language test. Students who are confident (less anxious – lower scores on the questionnaire) get higher scores on the test.

If the two tests were designed to test unrelated abilities – the first is a test of language ability, the second is a test of basketball skills – there would be no reason to expect students’ scores to be similar and a ‘no relationship’ picture in the centre of Figure 9 would be anticipated. Anything close to an ideal positive or negative relationship would give us reason to suspect that at least one of the tests was flawed.

On the other hand, if ‘no relationship’ was the picture that emerged when we compared TEA scores with EAT scores, this would be catastrophic for the test developers. The results on the TEA would tell us nothing at all about a student’s likely performance on the EAT test (and EAT scores would tell us nothing about likely performance on TEA). Some students do well on both, but just as many do well on TEA and poorly on EAT, others do well on EAT, but poorly on TEA. The scores appear to be scattered about at random.

If we found this result, even though the two tests were designed with the same construct in mind, the two tests could not possibly be measuring the same thing. Obviously we would not be able to use the results of tests that show this kind of relationship to draw any meaningful conclusions about students’ actual abilities. Trying to interpret the results would be like wearing two watches when at least one of them is faulty. When one shows five o’clock, the other shows three; when the first
shows six o’clock the other shows nine. Judging the actual time from this evidence could only be a matter of guesswork.

This is a matter of reliability. If different sources of evidence about assessee’s abilities support each other, giving a consistent picture, we can say that the evidence is reliable. Provided that it is also valid, it will make a sound basis for decisions about assessee. However, if the picture is vague or contradictory, it can never provide a good basis for decision making.

The same principles apply if, rather than comparing two tests, we compare scores awarded independently by two different raters (or by the same rater on different occasions). If both raters agree on which students are more and which less able, this suggests that both agree to a large extent about what they are measuring. If different raters order the assessee in very different ways, this is a clear sign that they are not able to use the scoring system effectively. Action is needed. What that action should be – more training for the raters, better rating scales or new raters – is a matter of judgement (or further investigation) according to the situation.

**Correlations**

So far, we have compared scores by displaying them on scatterplots (as graphs like those in Figure 9 are known). But a summary of the same information can be provided by a single number: the correlation coefficient.

Correlation coefficients can be calculated in different ways according to the nature of the data, but they range from -1 (the ideal negative relationship) to 1 (the ideal positive relationship). A correlation of 0 would signal that there was no relationship at all between two sets of scores. These figures are shown below the scatterplots in Figure 9.

![Figure 10](image)

The scatterplots in Figure 10 show how higher correlations reflect closer relationships between scores. We can plot a straight line (the regression line) that mimics a perfect positive relationship by passing as close as possible to both sets of scores. With a correlation of 0.9, most points are very close to the line: both tests seem to be measuring something very similar. If you knew a student’s score on one test, you could estimate what her score would be on the other and you would probably be very close to the truth.

The distance between each individual student’s scores and this line tells us something about how well that score fits with the overall picture we are building up. In the scatterplot in the middle of Figure 10, the correlation between two sets of scores is moderately high at 0.8 and the points generally fall close to the line. However, we can tell that there is something rather unexpected about Bela’s scores because his results fall a relatively long way from the line. His results do not fit with the picture of the close relationship between the tests that we have built up from other students’
results. As with the case of Kanya above, we need to look for a reason for the inconsistency between Bela’s results.

When the correlation is much lower (as it is on the right of Figure 10 – at 0.5), the overall picture is much harder to see. A lot of students have results that fall a long way from the line. It is much more difficult to pick out individual assesses with unusual results and much more difficult to see how the two sets of scores might be linked.

Correlations are helpful when evaluating the reliability of an assessment. If two forms of a test, designed to the same specifications, give very similar results – if the correlation between them is close to 1 – the evidence they provide would seem to be reliable.

**Internal consistency**

While comparisons between assessments – scores on different forms of a test; scores from the same assessment given on different occasions; scores awarded by different raters – offer insights into reliability, it is also important and usually more practical to consider the internal consistency of an assessment: the extent to which a single assessment produces a coherent picture of the assesses’ abilities.

If students like Carla have high scores and students like Luigi have low scores on Section A of the TEA test and the same is true on Sections B and C, the correlation between all parts of the test will be high: the evidence is consistent and the test as a whole is reliable. Internal consistency is often reported using a statistic called Cronbach’s alpha which is based on the mutual correlations between all of the items on an assessment.

As with the correlations between different tests, the ideal value for alpha, representing perfect reliability, is 1. Just as a low correlation between two assessments reveals that a student’s score on one will be difficult to predict from their score on another, so a low alpha suggests that we cannot be very sure that a student’s score on the test is a very accurate indication of how well he or she is really capable of performing. The evidence provided by the assessment is contradictory.

**True scores**

In reliability theory, all test scores are made up of two elements. One is called the ‘true score’. The other is ‘error’. Together they make up the ‘observed score’ or the score the assesssee was actually awarded. These concepts are explained in the following section.

How many steps does it take you to walk ten metres? Now try estimating the distance in metres from one point to another (e.g. the length of one city block) by counting how many steps you take to walk it. Try it just once and you can be pretty sure that you will get the measurement wrong. There are a number of reasons for this: people do not walk in a completely straight line, each footstep may be slightly longer or shorter than one metre. Try it again and you will probably get a different result. If you keep trying, sometimes you will overestimate the distance, sometimes you will underestimate it. However, you will find that the results seem to cluster around a certain point.

If you carried on pacing up and down for long enough and made a graph showing how often you came up with each figure, you would find that your results formed a normal distribution. The mean of your measurements is probably quite an accurate representation of the actual distance. However on each occasion there is a certain amount of random variation – error – in the estimate.

The standard deviation tells us how consistent you have been. Imagine you worked out the actual distance to be around one hundred metres and the standard deviation of your estimates was just
one metre: you were very consistent in your measurements. If you were to pace out the length of the next block, there’s a very good chance that you’d be within 1 or 2% of the correct distance. If you found that it was about 90 metres long, you might feel quite confident in saying that the first block was longer than the second. On the other hand, if the standard deviation of your first set of measurements was 10 metres, it’s quite likely that your next measurement would be 10% or even 20% more away from the correct distance. You couldn’t feel very confident about comparing the two distances because there is so much random variation – measurement error – involved.

As explained in Exploring Language Testing and Assessment (Chapter 4), in language assessment we don’t have any objective way of knowing the equivalent of the ‘correct distance’ – an assessee’s actual level of knowledge, skill or ability. Like pacing the block again and again, if we could get an assessee to repeat a test an infinite number of times, we could find their average score and take that as our best estimate of their actual ability to perform on the test – their ‘true score’.

In reality, most language learners are strangely reluctant to take the same assessment more than once, let alone an infinite number of times. This means we can never actually find the true score, we can only estimate how close our measurements might be to the true score.

If the test is very reliable – if all of the parts of the test give a consistent picture of the assessee’s abilities – we can conclude that most of their observed scores must be pretty close to their true score. If the test is unreliable – different parts of the test provide contradictory evidence – a lot of the scores must be a long way from the true score: factors other than what the test measures must be playing a big part in the observed scores that we obtain.

**Standard error of measurement**

We can use the reliability and dispersion of the scores on a test to calculate a standard error of measurement (SEM) (refer to the further reading sources for the calculations involved). The SEM is reported as a number of points on the test scale and provides an estimate of the error associated with individual scores. It is an estimate of what the standard deviation of the observed scores would be around the true score if each assessee were to take the assessment hundreds of times. If the observed scores were often a long way from the true score, the SEM would be large. If the observed scores were generally very close to each other, the SEM would be small. Based on what we know about the properties of the normal distribution, we can assume that 68% of assessee’s scores on the test would be within one standard error of their (unknowable) true score and that 95% would be within two standard errors.

Often the SEM is presented as a kind of ‘confidence interval’ around a test score. We know that Carla scored 74 on the TEA test. Let’s say we find that the SEM is 3 points. Using what we know about the normal distribution, this would suggest (with 68% confidence) that her true score lies within three points of 74 (71 to 77). We can be very confident (95% confident) that it is somewhere between 68 and 80 (although we have no way of knowing if it is higher or lower than her observed score).

Luigi scored 35 so we may suppose his true score is somewhere in the range of 29 to 41. It seems that Carla would almost certainly perform better than Luigi regardless of the element of error in the scores. On the other hand, we cannot be so sure that Ben, who scored 77, is really more able than Carla. He may just have been luckier than Carla on this occasion. On the other hand it is quite
possible that the difference between their true scores may be greater than the difference between their observed scores. If the pass mark on the test is 75, we would have to admit that we can’t be very confident that Ben really should pass the test and that Carla should fail. Ideally, we want to have more evidence about their abilities before choosing between them.

The terms ‘true score’ and ‘error free measurement’ may be a little misleading. Although the words seem to imply a theoretically perfect representation of the assessee’s abilities – a score that reflects only what we really want to measure – this is not really the case. In fact the true score only reflects the result we would obtain if the assessment were perfectly consistent: what we are measuring may not be the same thing as what we intend to measure. There is no guarantee that an assessment with a very low SEM is actually measuring the right abilities – that the evidence it provides is all relevant to the purpose of the assessment and that it is sufficient for the decisions we wish to take about the assessee. In fact, given the complexity of language use, very high reliability may be a sign that only a rather limited range of abilities is being assessed. Convincing evidence of validity is also needed before we can feel confident about using scores even from a very reliable test to help us take important decisions.

*What is an acceptable level of reliability?*

Approaches to estimating reliability are rather different according to whether the test is designed for norm referenced or criterion referenced interpretations. Alternative indices and ways of calculating and interpreting them are covered in the further reading sources suggested below.

Although rules of thumb for acceptable levels of reliability have been put forward (Lado 1961 suggested 0.90 to 0.99 would be reasonable for a tests of grammar, vocabulary and reading comprehension; 0.80 to 0.89 for tests of listening and 0.70 to 0.79 for speaking tests), the SEM is really a more useful indicator. Unlike an overall reliability figure, the SEM relates to the accuracy of an individual assessee’s score.

Of course, the SEM should be as small as possible, but as Lado’s recommendations make clear, reliability is more difficult to achieve for some kinds of assessment than for others. It should also be kept in mind that the SEM is variable. Scores that are close to the mean are less prone to error than those that are further away. If test results are used for making pass/fail decisions, it will be important to know the SEM at the pass mark. If several grades or bands are awarded, an SEM should be reported for each one.
**Task 3**

Investigate three international tests on the internet. How are scores reported? Do any use standardised scores?

Is any data provided on mean scores and standard deviations?

Is the reliability of the test reported? Is a standard error of measurement given?

How are these statistics explained?

---

**Item analysis**

I have said that statistics can function like alarm bells or flags to alert us to issues that we may need to address. Scores can show us where additional instruction should be targeted. Reliability statistics show us where we need to be most cautious in interpreting results and where assessments may not be functioning consistently.

In order to improve the quality of our assessments – to make them more reliable and more valid – we need to be able to diagnose where individual items may be contributing to the inconsistency of results. If we are able to identify problematic items, we can revise or replace them to improve the quality of the assessment.

According to whether the assessment is used for norm referenced or criterion referenced interpretations, slightly different qualities in the assessment material will be important. In criterion referenced assessments, the most useful items are very sensitive as detectors of learning. In other words, an ideal item is one that every student gets wrong before studying a unit, but that everyone gets right after studying the unit.

Items intended for use on criterion referenced tests should be trialled both on a pre-test and on a post-test or on forms of the test administered to people who are known to have the abilities in question (masters) and those who are known not to have these abilities (non-masters). We can compute a simple difference index by comparing the proportion of assesses giving correct responses on each occasion or in each group.

The proportion of assesses responding correctly to an item – also known as the item facility or p-value – is represented by a number that can range from 0 to 1. An item facility of 1.0 shows that all assesses responded correctly to the item; 0.5 shows that half of the assesses responded correctly; 0.0 that none responded correctly. In choosing items for criterion referenced assessments, we are interested in the difference between the item facility value for the group of masters and the facility value for the group of non-masters.
In Figure 11, on a pre-test (or test for non-masters), 65 out of 500 assesses responded correctly (a proportion of 0.13) to item 2. On the post-test (the test for masters), 430 (0.86) gave the correct response. The difference index is 0.86 – 0.13 or 0.73. This item seems to be highly sensitive to instruction. If there is no difference in the proportion of students who get an item right before and after instruction (or if more got it right before instruction), there must be a problem that we need to investigate. Perhaps the item itself is faulty or perhaps the learning target is unrealistic.

**Difference index**

Brown and Hudson (2002) suggest that a difference index of 0.2 or higher would make an item worth considering for inclusion on a criterion referenced test. Where master – non-master comparisons can be made, they cite Berk’s (1980) suggestion that good items are those with an item facility of 0.7 or higher for the group of known masters and 0.5 or less for the group of known non-masters. As it is not always practical to give exactly the same assessment before and after instruction, Brown (2005) also suggests a number of alternatives to the difference index that can be calculated on a single test administration.

In norm referenced assessment, an ideal item is one that allows us to discriminate efficiently between more able and less able assesses. Unlike on criterion referenced tests (where we need to know that every assessee has mastered the course content), items that almost everyone gets right are of little use on norm referenced tests because they do not differentiate efficiently between the assesses.

An item may not be particularly useful for a norm referenced test if its item facility is higher than around 0.8 (most people get it right) or lower than about 0.3 (most people get it wrong). An exception to this rule would be a placement test where a number of cut points need to be established. In this case it may be important to distinguish among students even within the highest scoring and lowest scoring ranges: who should go into the very highest ability class in the school and who into the second highest? Where multiple choice or other selected response items are used item facilities will be meaningless below the proportion of assesses that might answer correctly just by guessing. For items with three options, a facility of 0.33 is to be expected even if no assessee knows the correct answer.

We can assume that assesses who score well on the test as a whole generally have more ability than those who get low scores. A good item will probably be answered correctly by most of the high scoring students, but most of the low scoring students will get it wrong.
Item-total correlation 0.62  Item – total correlation 0.14  Item – total correlation 0.26

Figure 12

We can create a bar chart to show the proportion of assesses at different score levels who answer an item correctly (or the average score on items that are scored using a scale). The charts in Figure 12 show the extent to which performance on each item is consistent with performance on the rest of the test. On item 4, assesses with higher scores on the test were more likely to respond correctly to the item. Only one in twenty of the assesses who scored between 0 and 10 on the 50 item test gave a correct response. In contrast, almost all of those scoring between 41 and 50 answered this item correctly. Note that this is very much like the ‘ideal positive relationship’ that we saw in Figure 9.

Items 5 and 6 do not present such a healthy picture. Although on Item 5 a somewhat higher proportion of top scoring than low scoring assesses respond correctly, the difference is not substantial. The highest scoring assesses are only marginally more likely to give a correct response than those scoring between 11 and 20 points on the test. This item does not seem likely to help us to discriminate between the more and less able assesses.

Item 6 presents an unexpected pattern. As we would expect, those scoring between 21 and 30 are more likely to give the right answer than those scoring between 0 and 10 or between 11 and 20. However, contrary to expectations, those scoring between 31 and 50 are actually less likely to respond correctly than those in the 21 to 30 band. It appears that greater language ability may actually be encouraging more incorrect responses. A result like this would need to be investigated further. Is there some obvious reason why higher ability learners get it wrong? Should the item be rejected, or can changes be made that might make it more useful? See the example of the Listening test item in Exploring Language Assessment and Testing, Chapter 5.

Item-total correlation

Just as the correlation between two forms of a test indicates how consistent a picture they produce, so the correlation between the results on individual items and the results on the rest of the test can inform us about the contribution that each item makes to the reliability of the whole. Again, correlations can range from -1 to 1 and the higher the number the greater the consistency between that item and the rest of the test.

Any figures that are obtained need to be understood in context. If only a small number of assesses take the assessment (at least 30 and preferably over 100 sets of assessment scores are needed for correlations to be readily interpretable), or if the figures come from a trial involving assesses that are not very similar to those who will take the operational assessment, the numbers will need to be balanced with cautious judgement. However, correlations of 0.4 and above are generally considered
good and values between 0.3 and 0.4 acceptable. Low correlations of less than 0.2 are clearly problematic and the items affected should be investigated and probably rejected. Correlations between 0.2 and 0.3 are marginal and should be investigated for possible problems. For example, there may be something ambiguous about the correct answer, or the wording of the instructions may be confusing. Negative item-total correlations are usually a sign that the answer key is wrong.

**Distractor analysis**

Where selected response formats are used, it is important not only to consider the proportion of assessees choosing the correct response, but also to investigate the other (incorrect) alternative answers: the distractors. This often helps to reveal why a particular item may not have functioned well.

First, if very few assessees choose a distractor, it is not working as it should. Bachman (2004) suggests that as a ‘rule of thumb’ for multiple choice items each distractor should have a ‘facility’ value of at least 0.1: it should be chosen by 10% of the assessees. Where there are a large number of distractors, this proportion will necessarily be smaller, but each distractor should be chosen by a similar proportion of assessees. If one or more of the distractors appears very obviously incorrect to everyone, it will be much easier for assessees to guess the correct answer.

The item-total correlation for each distractor should be negative. This means that the assessees choosing the distractor in preference to the correct response tend to get lower scores on the test. If a distractor has a positive correlation with test scores it may also be a plausibly correct response. If the intended ‘correct’ response also has a negative correlation with scores, this is usually a sign that there is a mistake in the answer key.

**Statistics in decision making**

Statistics are not the only, or necessarily the best, tool for checking quality. Rigorous inspection of material and focussed discussion of results, even in the absence of any statistics, are very useful ways to improve quality. There can even be a danger that the pursuit of statistical indicators (such as a high alpha coefficient) can distort assessment. Setting rigid statistical targets can sometimes lead to poor decision making. The objective of any language assessment is not to obtain the best reliability estimate, but to obtain sufficient relevant evidence to support sensible decision making.

Even when speaking and writing skills are a core part of a language teaching programme, people sometimes argue that they should not be assessed because this cannot be done sufficiently reliably. This can lead in turn to unfortunate washback effects (see Chapter 4 of *Exploring Language Assessment and Testing*). Speaking and writing may be given less attention in the classroom because they do not appear on the test. Statistical indicators must be balanced with other considerations: speaking and writing must be assessed if they are a valued part of the curriculum. The challenge is to assess them in a way that reflects the programme objectives, but to do this as reliably as possible.

It follows that the assessment with the lowest standard error is not always the most suitable assessment for a particular purpose. In evaluating assessments, full account must be taken of all the qualities that go into making an assessment useful (*Exploring Language Assessment and Testing*, Chapter 4).
Statistics in Language Testing Research

Studies involving statistics are a mainstay of language testing research. Much of this research is directed at understanding the nature of relationships between the knowledge, skills and abilities being tested. The techniques used are not generally especially complex, but they are well beyond the scope of a basic guide. The interested reader is advised to begin with one of the books recommended below.

Further reading

Introductory books like Bachman (2004) and Brown (2005) offer guidance on additional resources that will be useful to those wishing to explore the language assessment research literature in greater depth. Bachman and Kunnan (2005) provides hands-on exercises for the reader to work through.


References


Ranks
If we rank the students according to their scores, we can easily see that Irma was the top scorer with 90 points and Gladys was the lowest with 13.

1st Irma 90
2nd Ben 77
3rd Carla 74
4th James 55
5th Amy 53
6th Freddy 34
7th Emma 30
8th David 25
9th Harry 23
10th Gladys 13

Means
Adding up all the scores gives a total of 474 points. There were 10 people who took the test so the mean is 474 divided by ten or 47.4 points.