The Standard Error of Measurement and Confidence Intervals

Barry Johnson, Jacky Ridsdale, Anwen Jones, Judy Capener

Introduction

This article represents a joint publication from PATOSS and Dyslexia Action on the principles of confidence intervals and best practice in their use, interpretation and reporting in SpLD assessment. While relevant for all assessors, it will be particularly helpful for those applying for assessment practising certificates or who must prepare a diagnostic report for submission as part-evidence of their professional competence.

What is a 'confidence interval'?

When specialist teachers and psychologists interpret test scores it is vital they consider and fully understand confidence intervals. If they do not they may draw invalid conclusions as to the presence or otherwise of specific learning difficulties.

The need for a confidence interval reflects 'Classical Test Theory' which states that there is *always* error in tests. Each individual diagnostic test has a predictable amount of error. This is known as the 'standard error of measurement' **(SE_m).** Confidence intervals (sometimes also referred to as confidence 'ranges' or 'bands') are derived using **SE_m** data, and thus perform the role of acknowledging this test error.

A confidence interval is a range of scores in which we can be confident that a person's 'true' score lies. The *degree* of confidence required is decided by the assessor in advance - 68%, 85%, 90% or 95% levels of confidence are commonly quoted - reflecting different degrees of certainty that the true score lies within the selected range.

Note that the *higher* the degree of confidence required, the *wider* will be the range of possible scores. So, a 68% confidence interval, where you are only confident that approximately 2 times out of 3 the true score lies in the quoted range, will be *narrower* than a 95% confidence interval, where you are confident that approximately 19 times out of 20 it does. For example, consider the different confidence intervals around a standard score of 90 on a test with four sub-tests:

Subtest	StandardScore	85% Confidence Interval	90% Confidence Interval	95% Confidence Interval
Word Reading	90	84 – 97	83 – 97	82 – 99
Sentence Comprehension	90	84 – 97	83 – 98	81 – 100
Spelling	90	83 – 98	82 – 99	80 – 101
Maths	90	82 – 100	81 – 101	79 – 103

Note how the width of the confidence ranges becomes larger as the selected level of confidence increases.

Returning to the **SE_m,** which is used to calculate confidence intervals, the formula to calculate it is:

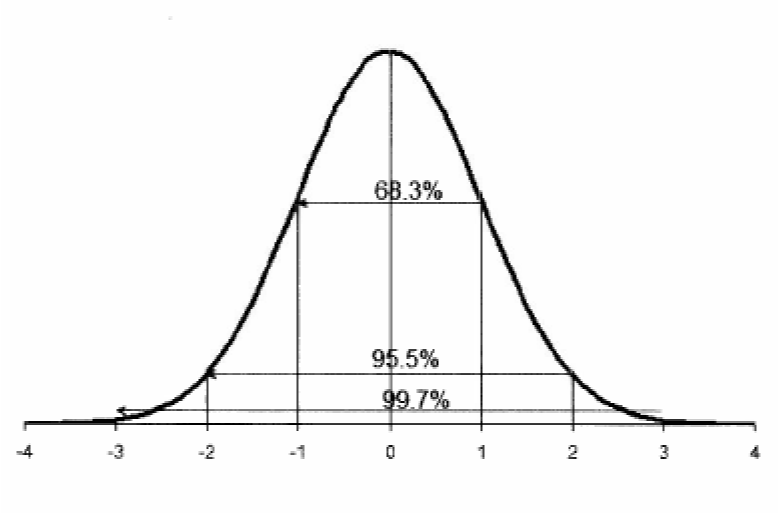
$$SE_m = SD\sqrt{1 - r}$$

Where for the test concerned, *SD* = Standard Deviation, and *r* = Reliability Coefficient.
Many tests provide the **SE$_m$** information within the reliability section of the test manual.

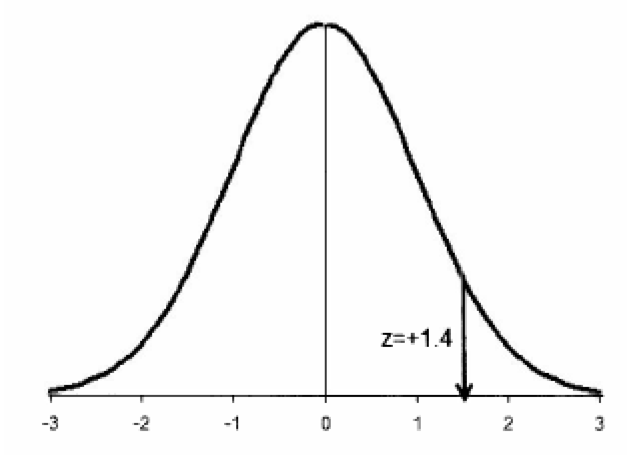
As most tests in common use have a standard deviation of 15, it is really the *reliability* of the test that determines the size of the **SE$_m$.** Thus, the *higher* the reliability, the *smaller* the standard error of measurement. The significant impact reliability has on the confidence interval will be shown later.

But how do we obtain the varying measures of confidence intervals from the **SE$_m$?** Firstly, we make the assumption that test measurement errors are normally distributed. In other words, if we took hundreds of measurements on the same test from the same person, we assume the scatter of errors of all these measurements (i.e. how far away all the scores were from the "true" score) would reflect a standard normal distribution.

A standard, normal distribution curve is symmetrical around a central point, and split into standard deviation units. The curve has a range of four standard deviations to each side of its central point. The number of standard deviations is often reduced to *three* either side of the central point for practical reasons and which reflects the very low number of scores covered by the extreme tails of the curve. Within each of these standard deviations, there is *always* a fixed proportion of area under the curve or, in this case, a fixed proportion of the possible range of scores.



The measurement of how far away from the central point a particular score lies is given in standard deviation units and is referred to as the z score. Thus a *z* score is another way of talking about standard deviations. It is useful as it acknowledges whether a score is above or below the central point by being either positive or negative. For example, in the figure below, a z score of +1.4 is marked.



How can we calculate a confidence interval?

The relevance of the discussion of **SE$_m$** and *z* scores can be seen when we consider the formula for a confidence interval:

$$CI = score \pm z(SE_m)$$

Z	= 1	68% level of confidence
Z	= 1.44	85% level of confidence
Z	= 1.65	90% level of confidence
Z	= 1.96	95% level of confidence
Z	= 2.58	99% level of confidence

Note it is often the case that **z** score values are rounded to the nearest whole number: for the 95% level of confidence, the **z** score is often taken as 2, for the 99% level of confidence, the **z** score is often taken as **3.**

Thus, to calculate a confidence interval around an obtained score we proceed along the following steps:

Step 1	Locate or calculate test's **SE$_m$** (see reliability section of test manuals)
Step 2	Select the confidence interval to use (85%? 95%? etc)
Step 3	Decide the *z* value that relates to your chosen level of confidence (use table above)
Step 4	Note the person's obtained test score
Step 5	Apply confidence interval formula

Example: Consider Test B where an individual obtained a score of 110

Step 1	Test B **SE$_m$** = 5
Step 2	95% confidence interval selected
Step 3	*z* = 2 (use table above and rounding as noted)
Step 4	Obtained test score = 110
Step 5	Lower limit of confidence interval

	= score - (z x $\mathbf{SE_m}$) = 110 - (2 X 5) = 100 Upper limit of confidence interval = score + (z x $\mathbf{SE_m}$) = 110 + (2 X 5) = 120 95% Confidence Interval is therefore 100 - 120

The impact of test reliability on confidence intervals

Small differences in reliability can lead to very significant differences in $\mathbf{SE_m}$ and thus significantly affect the confidence interval, as the following example shows.

Consider two tests each with a standard deviation of 15 but different reliability coefficients. Following the procedures above it follows that for a test score of, say standard score 106, the confidence intervals will be:

	Standard Score	Test 1
		Reliability = 0.75
		SEm = 7.5
			Test 2
			Reliability = 0.96
			SEm = 3

68% confidence interval =	106	98 – 114	103 – 109
90% confidence interval =	106	94 – 118	101 – 111
95% confidence interval =	106	91 – 121	100 - 112

The importance of using tests with high reliability coefficients is clearly portrayed in the table above, as we can see that the confidence intervals for Test 2, the test with higher reliability, are much narrower than for Test 1, which has lower reliability.

One criterion of a good test is if it provides confidence intervals alongside individual test scores, reflecting contemporary, agreed quality assurance standards in the industry. Test manuals which do not provide either confidence intervals, reliability data or $\mathbf{SE_m}$ information should be regarded as relatively unsafe and deserving of some caution.

How does one decide on which confidence interval to use?

Some test manuals offer a choice of confidence intervals, others only one, and a few offer none. Assessors have the responsibility to pre-select a confidence interval appropriate and safe for the decisions required for the individual client. It is considered inappropriate to decide **retrospectively** the level of confidence having obtained scores and derived confidence ranges.

A benchmark norm in statistics is to use the 95% level of confidence. This is widely used by psychologists and specialist teachers and is the recommended approach.

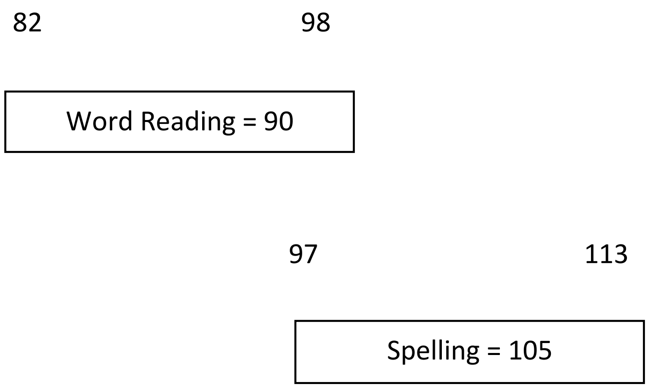
How can confidence intervals be used for the diagnostic process?

Confidence intervals enable the diagnostic teacher or psychologist to appreciate that a client's score on a test on any one day should not be regarded as their 'true' score. The obtained score is seen as lying within a confidence range, which can then be used to determine possible meaningful differences between subtest scores from the same test battery. This approach can also be used to compare scores in tests which have been co-normed. [Note: This approach cannot be used for tests that have been normed on different sample populations.] By taking into account the confidence interval the diagnostic teacher or psychologist will avoid 'over-egging the pudding' when appraising differences. The following table of hypothetical subtest scores is given as an example.

Subtest	Standard Score	95% Confidence Interval
Word Reading	90	82 – 98
Sentence Comprehension	75	67 – 83
Spelling	105	97 – 113

Without the column of confidence intervals for the test scores, one may have been tempted to conclude that the standard score of 90 for Word Reading is significantly below the standard score of 105 for Spelling. In fact, *for the pre-selected level of confidence,* the two confidence ranges overlap and therefore this conclusion cannot be validly made.

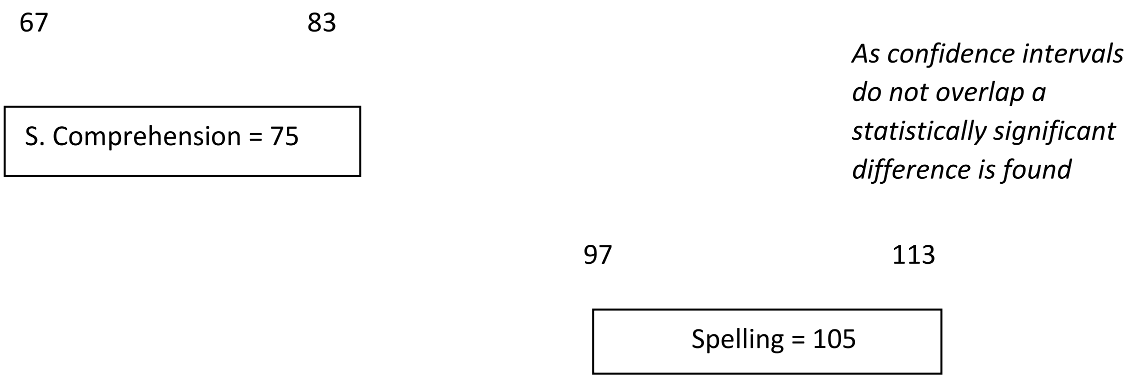
This can be seen more clearly if you were to present the confidence intervals in a visual display or graph. Here you see that the 2 confidence ranges overlap.



As confidence intervals overlap no statistically significant difference is found

It is important to note that there are more sophisticated means of comparing scores which take into account the statistical phenomenon of *regression to the mean.* This area is beyond the limits of this article, but given this factor assessors should be cautious when interpreting these overlaps when they occur at the extremes of the confidence range as in this example.

However, when the upper edge of one confidence band *does not* overlap with the lower edge of another, this firmly suggests a statistically significant discrepancy between the two scores does exist. Take the comparison between the Sentence Comprehension and Spelling performance in this example.



As confidence intervals do not overlap a statistically significant difference is found

Where a statistically significant difference is shown, this suggests a detectable, noticeable difference does exist, which did not occur by chance.

It is important to note that although a difference may be statistically significant it may not necessarily be very rare and assessors must also take this into consideration when coming to their conclusions. This is why some tests such as the WRIT and WRAT 5 provide tables for assessors to consider rarity (also known as ‘prevalence’ or ‘base rate’). It is normal for all people to have variation in their profile of test scores; the clinical relevance of the variability is dictated by the degree of the differences between test scores.

At every stage assessors should be aware that familiarity with statistical analyses, including the use of confidence intervals, must be balanced with their knowledge of the qualitative aspects of assessment. It is this balance of qualitative evidence with strong quantitative data which marks out the most useful assessments.

How can confidence intervals be reported to the lay reader?

Diagnostic teachers and psychologists should always strive to explain their results to the lay reader using consistent, clear and understandable language, avoiding the use of complex words and terms. Thus, the demands for an easily accessible report must be balanced against the need to provide full information.

It is advisable for the written report to follow the SpLD Assessment Standards Committee (SASC) Diagnostic Assessment Report Format and provide an overview, main body, confirmation of diagnostic decision and appendices, including an explanation of statistical terms and a summary table of test results, including confidence intervals. The statistical appendix enables the professional reader to understand the report writer's rationale as well as check the accuracy and interpretation of the statistical findings.

However, professionals must apply their judgement in the main body of the report. It is important that lay readers understand the variability of scores but equally the report should not be overburdened with complex statistical terms and data. Decisions need to be based on individual circumstances and there are many journalistic styles and options from which to choose. It may be most helpful to explain in the Appendices how the use of confidence intervals might be relevant to a conclusion by identifying any significant differences found.

Are confidence intervals the same as range descriptors?

Confidence intervals are **not** the same as range descriptors. Range descriptors are qualitative terms used by test companies to describe band ranges of scores in which scores can be placed in categories such as *below-average, average, superior,* etc. The subtleties of using these range descriptors will be addressed in future articles.

What can I do if I have a test that does not give confidence intervals or $\mathbf{SE_m}$ information in its manual?

From knowing the test's reliability coefficient and the standard deviation, you can calculate the $\mathbf{SE_m}$ and the confidence intervals as explained above. If the test does not provide enough information, consider contacting the test publisher to ask for it. Otherwise, seriously consider obtaining a better test.

Some tests such as the WRIT only give partial data on confidence ranges but do give tables to help with sub-test score comparisons.

Why do my calculations of confidence ranges sometimes differ slightly to those reported in tests manuals?

Test companies use a range of sophisticated statistical techniques when calculating confidence intervals, although these are still based on the concept of $\mathbf{SE_m}$. The most important difference is consideration of the phenomenon of *regression to the mean.* Briefly, this means that if a first test score lies towards one extreme end of the normal distribution curve, there is more chance of the second score being nearer the mean than the first - there is a tendency for it to be 'pulled' towards the mean. So, when calculating the confidence interval for a score, particularly when it is an extreme score, it is argued that it is better to place the confidence interval not around the **observed** score, but around an **estimated** score that takes into account the impact of regression to the mean. As the estimated score will not be the same as the observed score, then the confidence interval will be different. Note that when tests have very high reliability coefficients, the differences between these ways of measuring are not that great. Thus, where manuals provide a confidence interval at a level selected by you, it is this data that should be reported.

It is important to note that this article cannot address in full the totality of statistical terms that are concerned with the understanding and interpretation of confidence intervals. The interested reader may wish to explore further terms such as the *standard error of estimation* and *simple regression to the mean* or attend a suitable SASC approved or BPS CCET Level 'A' course. Also, the reader should note that many test companies use Item Response Theory and its related statistical techniques to complement Classical Test Theory when constructing confidence intervals. This causes additional variation in the various obtained measures when referring to confidence intervals.

Main Points

- Confidence intervals are important when interpreting scores on psychometric tests
- Good test manuals provide information on $\mathbf{SE_m}$ and confidence intervals but the assessor can calculate confidence intervals when standard deviations and reliability coefficients are available
- Test reliability has a significant impact on confidence intervals - tests with high reliability coefficients should be preferred
- The assessor decides in advance on the level of confidence required when interpreting and comparing scores - the 95% confidence interval is most widely used

- Where confidence intervals of subtests within a test battery or of co-normed tests do not overlap, statistically significant differences are evident
- Assessors must always bear in mind qualitative aspects when drawing diagnostic conclusions, alongside statistical data
- Assessment reports should always contain confidence intervals in their Appendix in order for the reports to be critically examined
- The assessor needs to evolve styles of report writing that promote accessibility for the lay reader but which also acknowledge the error that is always present in test scores.

Original Authors:

Dr. Barry Johnson Principal Educational Psychologist and Head of Assessment Services in Dyslexia Action. He is also a BPS CCET Level 'A' assessor [retired].

Jacky Ridsdale is Principal Lecturing Psychologist in Dyslexia Action, a Local Authority Educational Psychologist in Sheffield and a BPS CCET Level 'A' Course Verifier and Assessor.

Anwen Jones is a Specialist Teacher Assessor, former leader of Specialist Teacher Training courses.

Judy Capener is the Patoss Board member. She regularly leads professional training and continues to work in private practice in assessment and teaching.

This article has been updated in 2020 in conjunction with PATOSS and can be used with their permission.

References

Glutting, J, Adams, W, & Sheslow, D (2000). Wide Range Intelligence Test. Wide Range, Inc: Wilmington: USA

SpLD Test Evaluation Committee (2016) DSA: STEC DfES Guidelines 2016 [Online]. Available from: http://www.sasc.org.uk/SASCDocuments/REVISED%20guidelines-March%202016%20a.pdf [Accessed 20.05.2020].

Wilkinson, GS and Robertson, GJ (2017) Wide Range Achievement Test, Fifth Edition (WRAT5), Bloomington, MN: Pearson Clinical