How Much Difference Does It Make?

Notes on Understanding, Using, and Calculating Effect Sizes for Schools

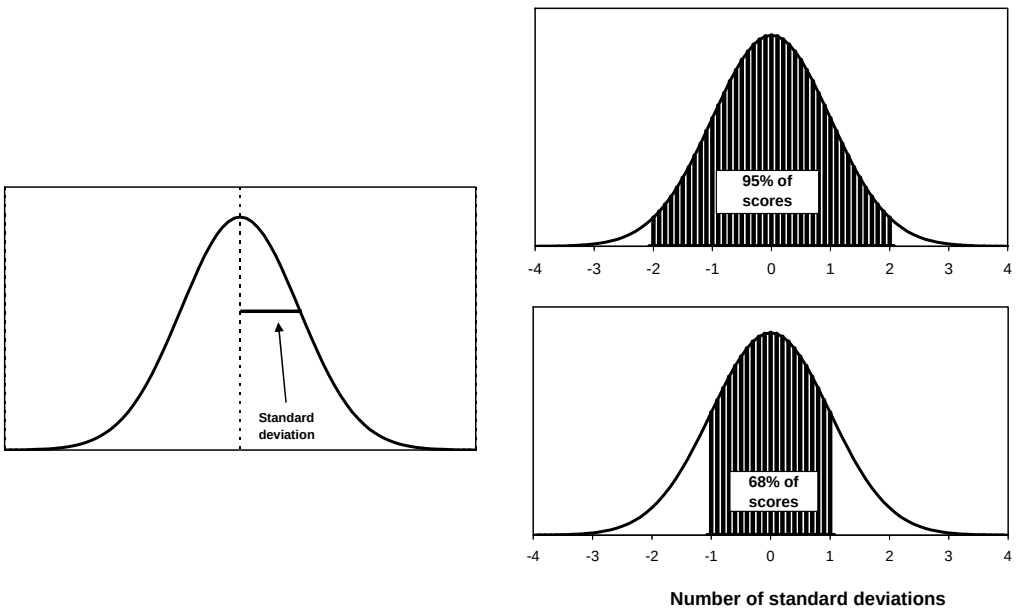
Ian Schagen, Research Division, Ministry of Education
Edith Hodgen, NZCER

Introduction

Suppose you tested a class in a topic and then gave them some kind of learning experience before testing them again, and on average their scores increased by 12 points. Another teacher in another school using a different test and another learning experience found a rise in scores of 25 points on average. How would you try to judge which was the better learning experience, in terms of improvement in scores?

Well, you can't just compare the changes in scores because they're based on totally different tests. Let's say the first test was out of 30 and the second out of 100 – even that doesn't help us because we don't know the spread of scores, or any way of mapping the results on one test into results on the other. It's as if every time we drove a car the speed came up in different units: kilometres per hour, then feet per second, then poles per fortnight. Not very useful.

One of the important aspects of any test is the amount of "spread" in the scores it usually produces, and a conventional way of measuring this is the "standard deviation" (often called SD for short). Many test scores have a hump- or lump- or bell-shaped distribution, with most students scoring in the middle, and fewer scoring very high or low. The theoretical distribution usually known as the "normal distribution" often describes test scores well. This diagram shows what the standard deviation looks like for an idealised test with a "bell-shaped" or normal distribution of scores.



When scores have a distribution like this, 68 percent of the scores lie within one standard deviation of the mean, and 95 percent lie within two standard deviations of the mean. Almost all scores lie within three standard deviations of the mean.

This standard deviation measure is a good way of comparing the spreads of different tests and hence getting a direct comparison of what are sometimes called "change scores". A change score is the difference between two test scores, usually for the same kind of test taken at different times. A change score is a way to measure progress.

There are actually two ways of getting a new measure from the test scores; one that is easier to compare in a meaningful way:

1. "Standardise" each test to have the same mean and standard deviation, so that you can compare score changes directly. For example, "IQ" tests tend to all have mean 100 and standard deviation 15; international studies (such as PISA and TIMSS) go for mean 500 and standard deviation 100.
2. Divide the change score, or difference between scores over time, $T_2 - T_1$, for each test by the standard deviation to get a fraction which is independent of the test used – we shall call this fraction an "effect size".

In this paper we focus on the second approach and try to show how to calculate, use and understand effect sizes in a variety of contexts. By using effect sizes we should be able to do the following:

- investigate differences between groups of students on a common scale (like using kilometres/hour all the time)
- see how much change a particular teaching approach makes, again on a common scale
- compare the effects of different approaches in different schools and classrooms
- know about the uncertainty in our estimates of differences or changes, and whether these are likely to be real or spurious.

What are standard deviations and effect sizes?

- The standard deviation is a measure of the average spread of scores about the mean (average) score; almost all scores lie within three standard deviations of the mean.
- An effect size is a measure that is independent of the original units of measurement; it can be a useful way to measure how much effect a treatment or intervention had.

Back to our example. Let's assume we know the following and we'll worry about how to get the standard deviation values later:

Class A: Test standard deviation = 10; average change in scores = 12; effect size = 1.2.
Class B: Test standard deviation = 30; average change in scores = 25; effect size = 0.83.

From these results we might be able to assume that there has been more progress in Class A than in Class B – but how do we know that this apparent difference is real, and not just due to random variations in the data?

So far we have introduced effect sizes and shown how they can be handy ways of comparing differences across different measuring instruments, but this now raises a number of questions, including:

- How do we estimate the standard deviation of the test, to divide the change score by?
- What other comparisons can we do using effect sizes?
- How do we estimate the uncertainty in our effect size calculations?
- How do we know that differences between effect sizes are real?
- How big should an effect size be to be "educationally meaningful"?
- What are the cautions and caveats in using effect sizes?
- How easy is it to calculate an effect size for New Zealand standardised tests?

Why use effect sizes?

- To compare progress over time on the same test (most common use).
- To compare results measured on different tests.
- To compare different groups doing the same test (least common use).

Getting a standard deviation

If we have a bunch of data and want to estimate the standard deviation, then the easiest way is probably to put it into a spreadsheet and use the internal functions to do it for you. If you want to calculate it by hand, here is how to do it:

1. Calculate the mean of the data by adding up all the values and dividing by the number of cases.
2. Subtract the mean from each value to get a "deviation" (positive or negative).
3. Square these deviations and add them all up.
4. Divide the result by the number of cases minus 1.
5. Take the square root to get the standard deviation.

Here is a worked example with the following values:

10, 13, 19, 24, 6, 23, 15, 18, 22, 17.

1. Mean = 167/10 = 16.7.
2. Deviations: -6.7, -3.7, 2.3, 7.3, -10.7, 6.3, -1.7, 1.3, 5.3, 0.3.
3. Squared deviations: 44.89, 13.69, 5.29, 53.29, 114.49, 39.69, 2.89, 1.69, 28.09, 0.09. Sum of these = 304.1.
4. Divide by 10–1 = 9: 33.79.
5. Square root: 5.81.

Therefore the standard deviation is estimated as 5.81. However, if we tested a different bunch of 10 students with the same test we would undoubtedly get a different estimate of standard deviation, and this means that estimating it in this way is not ideal. If the value we're using to standardise our results depends on the exact sample of students we use, this means our effect size measure has an extra element of variability which needs to be taken into account.

Another issue arises when we test and retest students. Which standard deviation do we use: the pre-test one, the post-test one, or some kind of "pooled" standard deviation? If we use the pre-test, then it may be that all students start from the same low state of understanding and the standard deviation is quite small (or even zero) – this will grossly inflate our effect size calculation. The same might happen with the post-test, if we've brought everyone up to the same level. The "pooled" standard deviation is basically an average of the two, but this might also suffer from the same issues.

A better option is to use a value which is fixed for every different "outing" of the same test and which we can use regardless of which particular group of students is tested. If the test has been standardised on a large sample, then there should be data available on its overall standard deviation and this is the value we can use. If it's one we've constructed ourselves then we may need to wait for data on a fair few students to become available before calculating a standard deviation to be used for all effect size calculations.

Another option is to cheat. Suppose we have created a test which is designed to be appropriate over a range of abilities, with an average score we expect to be about 50 percent. We also expect about 95 percent of students to get scores between about 10 percent and 90 percent. The normal "bell-shaped" curve (see diagram above) has 95 percent of its values between about plus or minus twice the standard deviation from the mean.

So if 90 – 10 = 4 x standard deviation, then estimate the standard deviation = 20.

If we use 20 as the standard deviation for all effect size calculations, regardless of the group tested, then we have the merits of consistency and no worries about how to estimate this. We can check our assumption once we've collected enough data, and modify the value if required.

Another example: Suppose we monitored students on a rating scale, from 0 (knows nothing) to 8 (totally proficient). Then we might say that the nominal standard deviation was around 8/4 = 2.0, and use this value to compute effect sizes for all changes monitored using this scale.

Measuring variability in test scores

- If possible, use the published standard deviation for a standardised test.
- For a test where there are no published norms, calculate the standard deviation for each set of data, and state whether you chose to take:
  - the standard deviation for the first set of scores
  - the standard deviation for the second set of scores
  - a pooled value that lies between the two (closer to the value from the set of scores that has more students)
  - an estimate (or "nominal SD") from the expected highest and lowest scores for the middle 95 percent of the students:
  - SD = (highest estimate – lowest estimate)/4.

Possible comparisons using effect sizes

The following broad types of comparisons are possible using effect sizes:

- differences in scores between two different groups (e.g., boys and girls)
- changes in scores for the same group of students measured twice
- relationships between different factors and scores, all considered together.

In the first type of comparison (between-group differences) we calculate an effect size by taking the difference in mean scores between the two groups and dividing that by the nominal standard deviation.

The second type of comparison (change scores) is what we have considered in the preceding section. The effect size is simply computed as the average change in scores divided by the nominal or assumed standard deviation.

Basically, the calculation is the same whatever we're doing: take a difference in mean scores and divide by a standard deviation.

The third type of comparison is more complex, but can be important. For example, suppose we have a difference between boys and girls, and also a difference between those who have done some homework and those who have not. There may be a relationship between these two groups, so that when we consider the data all together the effect size we get for each factor controlling for the other is smaller than it would be otherwise. Using a statistical technique such as regression we can estimate such "joint effect sizes" and compare the magnitude of the boy/girl difference with that of the homework/no homework distinction, each taking account of the other.[1]

[1] There are other, more appropriate, ways of measuring effect sizes in regression and related models, but these are outside the scope of this discussion.

Uncertainty in effect sizes

As with any statistical calculation, effect sizes are subject to uncertainty. If we repeated the exercise with a randomly different bunch of students we would get a different answer. The question is: How different? And how do we estimate the likely magnitude of the difference?

The term "standard error" (SE for short) is used to refer to the standard deviation in the likely error around an estimated value. Generally, 95 percent of the time the "true" value will be within plus or minus two SEs of the estimated value, and 68 percent of the time it will be within plus or minus one SE of the estimated value. If we assume the standard deviation of the underlying scores has been fixed in some way, then the SE of an effect size is just the SE in the difference of two means divided by the standard deviation.

When we are trying to measure the average test score, we expect that an average based on five students is less likely to be very near the true score for those students than an average based on 500 students would be. A single student having a very bad or good day could affect a five-student average quite a lot, but would have very little effect on a 500-student average. In the same way, the uncertainty in estimates of effect size is much greater for small groups of students than it is for large ones. In fact, as we'll see later, the uncertainty in effect size can be well approximated just using the number of students involved.

The calculations work differently if we are dealing with two separate groups or with measurements at two points in time for the same group. Let us do an example calculation both ways, assuming an 8-point scale with nominal standard deviation 2.0.

Here are some data, for two groups A and B:

	A	B	Diff B–A
	1	3	2
	3	4	1
	4	5	1
	3	6	3
	2	4	2
	5	4	-1
	3	3	0
	4	7	3
	1	3	2
	5	6	1
Mean	**3.1**	**4.5**	**1.4**
SD (from data)	**1.45**	**1.43**	**1.26**
SE = SD/ $\sqrt{n}$	**0.46**	**0.45**	**0.40**

In scenario 1, A and B are two separate groups whose means we wish to compare. The effect size[2] is (4.5–.1)/2.0 = 0.7. The SE for the mean of group A is calculated from the standard deviation of the group A scores divided by the square root of the number of cases (10), giving the value 0.46. A similar calculation for group B yields a value of SE equal to 0.45.

To get the SE for the difference in group means we need to combine these two separate SEs, by squaring them, summing them, and then taking the square root.

This gives: SE of group mean difference = $\sqrt{(0.46^2 + 0.45^2)} = 0.64$.

Therefore, SE of effect size = SE of group mean difference/(nominal SD) = 0.64/2.0 = 0.32.

A 95 percent confidence interval for the effect size is therefore 0.70 ± 1.96 x 0.32 = 0.07 to 1.33.

In scenario 2, group B is just the same set of students as group A, but tested at a later point in time. In this case we are interested in the difference scores, in the last column of the table above. The mean is 1.4, with a standard deviation of 1.26 and SE 0.40 (= standard deviation divided by square root of number of cases). The estimated effect size is still 0.70, but now with a value of SE equal to 0.40/2.0 = 0.20. A 95 percent confidence interval for the effect size is therefore 0.70 ± 1.96 x 0.20 = 0.31 to 1.09.

If we look at the size of the two confidence intervals, why is the second (0.31 – 1.09) so much narrower than the first (0.07 – 1.33)? In scenario 1, we measured *different* students on the two occasions, so some of the differences in score will be due to differences between students, and some due to what happened between testing points. In scenario 2, we measured the *same* students on both occasions, so we expect the second scores to be relatively similar to the first, with the difference between scores being mainly due to what happened between testing points, which means that the effect size is measured with less error.

A simpler estimate of SEs

An even simpler way of estimating SE values makes use of the fact that we've kind of cancelled out the actual standard deviations in the above formulae so that all we need to know to calculate the standard error is the number of students. The simple formulae are:

- Two separate samples (scenario 1): SE = square root of (1 divided by number in first group + 1 divided by number in second group) = $\sqrt{(1/10 + 1/10)} = 0.45$.

[2] The value 2.0 in the formula is the "nominal" or assumed value for our scale. Instead of this, we could used an average or "pooled" standard deviation estimated from the data as 1.44. This would give higher estimates of effect size, but would change if we took a different sample of students.

- Same sample retested (scenario 2): SE = square root of (1 divided by number in sample) = √(1/10) = 0.32, assuming a moderate relationship between test scores (a correlation of $r = 0.5$).[3]

The main reason why these "quick" estimates are different from those calculated earlier is that we have previously divided by a nominal standard deviation of 2.0 rather than the "pooled estimate" of 1.44. Had we used the pooled estimate we would have had 0.64/1.44 = 0.44 and 0.4/1.44 = 0.28.

This method for quickly estimating standard errors can be quite useful for judging the likely uncertainty in effect size calculations for particular sample sizes.

Measuring variability in effect sizes

- If different groups of students did the two tests, use
  SE = square root of (1/(number in first group) + 1/(number in second group))

  or use Table 6.
- If the same students did the two tests, use
  SE = square root of (2*(1-r)/number of students)
  where r is the correlation between the first and second test scores

  or use Table 7.

To make calculating effect sizes and their confidence intervals easier, we have made some tables for the main standardised tests used in New Zealand (asTTle, STAR, and PAT), see Tables 1–7 on pages 15–23. These tables allow you to read off an approximate effect size for a test, given a mean difference or change score. They assume that the difference is over a year, and take into account the expected growth over that year (see the examples). Example 3 (p. 12) shows how the tables can be used if the scores are not measured a year apart.

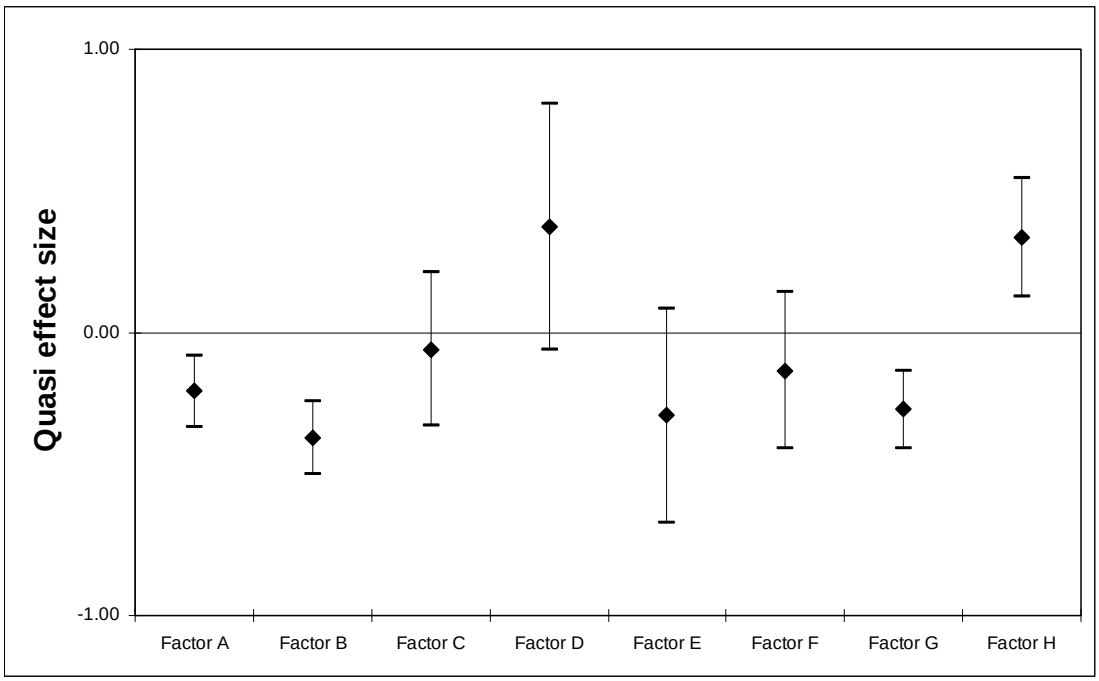
How do we know effect sizes are real?

This is equivalent to asking if the results are "statistically significant" – could we have got an effect size this big by random chance, even if there was really no difference between the groups or real change over time? Usually we take a probability of 5 percent or less as marking the point where we decide that a difference is real.

This is actually quite easy to do using the 95 percent confidence intervals calculated as in the above example. If the interval is all positive (or all negative) then the probability is less than 5 percent that it includes zero effect size, and we can conclude (with a fairly small chance of being wrong) that the effect size is really non-zero. A good way of displaying all this is graphically, especially if we are comparing effect sizes and their confidence intervals for

[3] If the correlation, r, between scores is known, then the formula is SE = square root of $(2(1 - r)/n)$, where n is the number of students.

different groups or different influencing factors. A "Star Wars" plot like the one below illustrates this.



In this kind of plot, the diamond represents the estimated effect size for each factor relative to the outcome, and the length of the bar represents the 95 percent confidence interval. If the bar cuts the zero line, then we can say the factor is not statistically significant. In the above plot, this is true for factors C, D, E and F. Factors A, B and G have significant negative relationships, while Factor H has a significant positive one. Although the effect size for H is lower than for D, the latter is not significant and we should be cautious about ascribing any relationship here at all, whereas we can be fairly confident that H really does have a positive relationship with the outcome. From what we saw above, the estimates for Factors D and E, with wide confidence intervals, would be based on far fewer test scores than those for Factors A, B, and G, with much narrow confidence intervals.

How big is big enough?

A frequent question is "How big should an effect size be to be educationally significant?" This is a bit like "How long is a bit of string?", as the answer depends a lot on circumstances and context. Some people have set up categories for effect sizes: e.g., below 0.2 is "small", around 0.4 is "medium" and above 0.6 is "large". But these can be misleading if taken too literally.

Suppose you teach a class a new topic, so that initially they pretty well all rate zero on your 8-point assessment scale. You would expect that most of them would reach at least the mid-point of the scale afterwards, with some doing even better. An effect size of 4.0/2.0 = 2.0 (mid-point of scale = 4; nominal standard deviation = 2) would not be an unreasonable expectation, but this would only be "large" within a restricted context. Similarly, if you took a small group with a limited attainment range and managed to raise their scores, you could get quite big effect sizes; but these might not be transferable to a larger population.

On the other hand, if you managed to raise the mathematics aptitude of the whole school population of New Zealand by an amount equivalent to an effect size of 0.1, this would raise our scores on international studies like PISA and TIMSS by 10 points – a result which would gain loud applause all round.[4]

Cautions, caveats, and Heffalump traps for the unwary

Effect sizes are a handy way of looking at data, but they are not a magic bullet, and should always lead to more questions and discussion. There may be circumstances which help to explain apparent differences in effect sizes – for example, one group of students might have had more teaching time, or a more intensive programme, than another. Looking for such apparent differences is one of the main reactions that effect sizes should lead to.

One thing to watch out for is "regression to the mean". This is particularly a problem when specific groups of individuals such as those with low (or very high) attainment are targeted for an intervention. If we take any group of individuals (class, school, nation) and test them, and then select the lowest attaining 10 percent, perform any kind of intervention we like, and then retest, we will normally find that the bottom 10 percent have improved relative to the rest. This is because there is a lot of random variation in individual performance and the bottom 10 percent on one occasion will not all be the bottom 10 percent on another occasion.

This is a serious problem with any evaluation which focuses on under- (or over-) performing students, however defined. It is essential that progress for such students be compared with equivalent students not receiving the intervention, and not with the progress of the whole population, or else misleading findings are extremely likely.

Whenever you calculate an effect size, make sure you also estimate a standard error and confidence interval. Then you will be aware of the uncertainty around the estimates and not be tempted to over-interpret the results.

Effect sizes are not absolute truth, and need to be assessed critically and with a full application of teacher professional judgement. However, if you believe some teaching initiative or programme is making a difference, then it should be possible to measure that difference. Effect sizes may be one way of quantifying the real differences experienced by your students.

[4] It would move New Zealand's average PIRLS score in 2006 from 532 to 542, above England and the USA.

Judging effect sizes

- The difference is "real" if the confidence interval does not include zero.
- The importance of the difference depends on the context.
- Groups consisting only of students with the highest or lowest test scores will almost always show regression to the mean (low scorers will show an increase; high scorers a decrease regardless of any intervention that has taken place).

How easy is it to calculate effect sizes for New Zealand standardised tests?

What you need to know before you can calculate an effect size is:

- the two sample means
- the expected growth for students in the relevant year levels
- the standard deviation.

The issues, really, are how to work out expected growth and a standard deviation. For standardised tests, it is best to use the published expected growth to correct any change in score to reflect only advances greater than expectation. Manuals or other reference material for the tests also give the standard deviation as published from the norming study, and this is the best value to use to calculate an effect size.

This means that, in fact, tables of effect sizes are easy to construct for a test, given the year level of the students (see Tables 1–5).

The examples below will walk you through both doing the actual calculations, and using tables to look up approximate values.

Example 1 (*PAT Maths, one year between tests*): If a group of Year 4 students achieved a PAT Maths score of 28.2 at the start of one year, and at the start of the next the same students achieved a score of 39.9 (in Year 5), they had a mean difference of 11.7, which is a little in advance of the overall mean difference of 8.5 at that year level (Table 1). To look up the effect size for this difference in Table 1, find the difference of 11.7 down the left side of the table, and Year 5 under PAT Mathematics across the top. The nearest difference down the left-hand side is 11.5, and the matching effect size is 0.23. Had the difference been 12.0, the effect size would have been 0.27, so if we calculated the effect size, rather than looking it up, it would probably have come out at around 0.24, which we could take as our estimate.

Alternatively, the effect size could be calculated directly using the data in the table. Our difference of 11.7 needs to be "deflated" by the expected growth (40.3 – 31.8 = 8.5) and then divided by the Year 5 standard deviation of 13.2. This gives an effect size of (11.7 – 8.5)/13.2 = 0.24.

Once we have an effect size, it is easy to add a confidence interval. Suppose that in the example above there were 57 Year 4 students and a year later, 64 students took the test in Year 5, and individual students were not matched (because, say, the school was one where there is a very high transience rate). The standard error of the effect size can be read off Table 6. Both samples are around 60 students, and matching standard error is 0.18. Had the samples been smaller, say both were of size about 50 (this was the nearest option in the table), the standard error would have been 0.20. If one sample was 60 and the other 50, the standard error would have been 0.19. So, taking all these options into account, 0.18 looks like a good estimate.

A 68 percent confidence interval for the effect size would be from 0.24 – 0.18 = 0.06 to 0.24 + 0.18 = 0.42, and a 95 percent confidence interval would be from about 0.24 – 2*0.18 = -0.12 to 0.60. Using the more stringent criterion, we cannot be sure that there was an effect.

What if we had more students? If we had 120 students in Year 4 and 140 in Year 5, the standard error would be somewhere between 0.12 and 0.14 (looking at the values for samples of 100 and 150, the nearest numbers in the table), so we can use 0.13. This would give a 68 percent confidence interval of 0.11 to 0.37, and a 95 percent confidence interval of -0.02 to 0.50, and we can still not be certain that there was an effect.

Example 2 (*PAT Reading, one year between tests*): A group of 60 Year 6 students had a mean PAT Reading comprehension score of 38.1 and when they were in Year 7 the mean score of the same students was 58.7. Their mean difference was 20.6, a great deal higher than the expected growth of 53.2 – 45.0 = 8.2. The effect size, from Table 1 is 0.98 or 0.99 (PAT Reading comprehension, Year 7, difference 20.5, which is the nearest to 20.6).

The standard error for this score is 0.08 from Table 7, assuming a correlation of 0.8 and sample of 60. This gives a 68 percent confidence interval of 0.99 – 0.08 = 0.91 to 0.99 + 0.08 = 1.07, and a 95 percent confidence interval of 0.99 – 2*0.08 = 0.86 to 1.15. Without doubt substantial progress was made in this case.

Example 3 (*asTTle Writing, two years between tests*): A school has had an intervention for two years. At the start, the 360 students in Year 4 at the school had an average asTTle Writing score of 390, and two years later the 410 students then in Year 6 had an average writing score of 521. Over the two years, they made a mean gain of 521 – 390 = 131. The tables are made for a single year's growth, so to use a table we need to "discount" the gain by the expected gain for the first year (the table will do the discounting for the second year).

The expected gain in asTTle Writing between Year 4 and Year 5 is 482 – 454 = 28 (from the top of Table 5), so our "discounted" gain is 131 – 28 = 103. In Table 5, the effect size for a mean difference of 103, for students now in Year 6, is between 0.78 and 0.83, so we can take a value of 0.81 (103 is a little closer to 105 than to 100).

The standard error of the effect size is between 0.07, 0.06, and 0.08 (using n_1 = 300 and 500 and n_2 = 300 and 500 in Table 6), so using 0.07 looks a good idea.

The confidence interval for the effect size is 0.81 ± 2*0.07 or 0.67 to 0.95.

We can say that the intervention appeared to be very effective.

Example 4 (*STAR, not quite one year between tests*): A group of 237 students had a mean STAR stanine score of 3.7 at the start of a year, and a score of 4.5 at the end of the year. STAR scores, and all other stanine scores, have a mean of 0 and standard deviation of 2. If a student progresses as expected, their stanine score will stay more or less the same over time.

Table 2 is provided for completeness, but effect sizes are very easily calculated for stanine scores (divide the mean difference by 2), and so long as the standardisation process was appropriate for each student's age and time of year, it doesn't matter how far apart in time the scores are (they do not need "discounting" for expected progress).

In this example:

Effect size = (4.5 – 3.7)/2 = 0.4 (or look up 4.5 – 3.7 = 0.8 in Table 2).

The standard error is about 0.03 as STAR tests tend to have a correlation of between 0.8 and 0.9, and the number of students is between 200 and 250, giving a confidence interval of 0.33 to 0.46. This would often be considered to indicate a moderate effect.

On the next page we present a brief summary of the ideas presented in this paper about effect sizes.

Summary of ideas on effect sizes

- Effect sizes are a useful device for comparing results on different measures, or over time, or between groups, on a scale which does not depend on the exact measure being used.
- Effect size measures are useful for comparing results on different tests (a comparison of two scores on the same standardised test is not made much more meaningful by using effect sizes).
- Effect sizes can be used to compare different groups of students, but are most often used to measure progress over time.
- Effect sizes measured at two different time points need to be "deflated" to account for expected progress (unless both measures are standardised against expected progress – for example, stanine scores).
- Published standard deviations should be used for standardised tests.
- For other tests, either an approximate SD can be guessed from the spread of scores, or the SD of the sample data can be calculated.
- Effect sizes should be quoted with a confidence interval.
- How the confidence interval is calculated depends on whether the same students were measured at the two time points (matched samples) or not.
- The confidence interval can be used to judge whether the effect is large enough to be considered unlikely to be a lucky chance (the interval should not include zero).
- Regression to the mean can produce a spuriously large effect size if the group of students being measured was selected as being the lowest performing 10 or 20 percent.
- The effect size measure discussed here is the most commonly used one, but is only really suited to comparing two sets of scores. There are other measures for more complicated comparisons.

Table 1 Effect sizes for PAT scores

The figures in the body of the table are the effect sizes for the achieved mean difference, at each year level and for each of the three tests, taking into account expected growth over **one year**. Where the difference is less than expected, the effect size is negative. The expected difference at each year level can be calculated from the mean PAT scores for two year levels. For example, from Year 4 to Year 5 in mathematics, the expected growth is 40.3 – 31.8 = 8.5.

		PAT Mathematics							PAT Reading Comprehension							PAT Reading Vocabulary						
Current year level		4	5	6	7	8	9	10	4	5	6	7	8	9	10	4	5	6	7	8	9	10
Mean PAT score		31.8	40.3	46.4	50.9	55.5	62.6	66.6	28.8	35.8	45.0	53.2	60.4	67.0	76.5	32.4	40.9	48.7	55.0	60.1	65.7	70.5
SD of PAT score		13.1	13.2	12.4	11.8	12.3	11.3	11.6	15.2	13.2	12.7	12.6	12.3	12.2	12.4	16.0	15.8	15.0	14.3	14.8	14.7	14.8
Difference between mean scores	7.0	-0.11	-0.11	0.07	0.21	0.20	-0.01	0.26	0.00	0.00	-0.17	-0.10	-0.02	0.03	-0.20	-0.09	-0.09	-0.05	0.05	0.13	0.10	0.15
	7.5	-0.08	-0.08	0.11	0.25	0.24	0.04	0.30	0.03	0.04	-0.13	-0.06	0.02	0.07	-0.16	-0.06	-0.06	-0.02	0.08	0.16	0.13	0.18
	8.0	-0.04	-0.04	0.15	0.30	0.28	0.08	0.34	0.07	0.08	-0.09	-0.02	0.07	0.11	-0.12	-0.03	-0.03	0.01	0.12	0.20	0.16	0.22
	8.5	0.00	0.00	0.19	0.34	0.32	0.12	0.39	0.10	0.11	-0.06	0.02	0.11	0.16	-0.08	0.00	0.00	0.05	0.15	0.23	0.20	0.25
	9.0	0.04	0.04	0.23	0.38	0.36	0.17	0.43	0.13	0.15	-0.02	0.06	0.15	0.20	-0.04	0.03	0.03	0.08	0.19	0.26	0.23	0.28
	9.5	0.08	0.08	0.27	0.42	0.40	0.21	0.47	0.16	0.19	0.02	0.10	0.19	0.24	0.00	0.06	0.06	0.11	0.22	0.30	0.27	0.32
	10.0	0.11	0.11	0.31	0.47	0.44	0.26	0.52	0.20	0.23	0.06	0.14	0.23	0.28	0.04	0.09	0.09	0.15	0.26	0.33	0.30	0.35
	10.5	0.15	0.15	0.35	0.51	0.48	0.30	0.56	0.23	0.27	0.10	0.18	0.27	0.32	0.08	0.13	0.13	0.18	0.29	0.36	0.33	0.39
	11.0	0.19	0.19	0.40	0.55	0.52	0.35	0.60	0.26	0.30	0.14	0.22	0.31	0.36	0.12	0.16	0.16	0.21	0.33	0.40	0.37	0.42
	11.5	0.23	0.23	0.44	0.59	0.56	0.39	0.65	0.30	0.34	0.18	0.26	0.35	0.40	0.16	0.19	0.19	0.25	0.36	0.43	0.40	0.45
	12.0	0.27	0.27	0.48	0.64	0.60	0.43	0.69	0.33	0.38	0.22	0.30	0.39	0.44	0.20	0.22	0.22	0.28	0.40	0.47	0.44	0.49
	12.5	0.30	0.30	0.52	0.68	0.64	0.48	0.73	0.36	0.42	0.26	0.34	0.43	0.48	0.24	0.25	0.25	0.31	0.43	0.50	0.47	0.52
	13.0	0.34	0.34	0.56	0.72	0.68	0.52	0.78	0.39	0.45	0.30	0.38	0.47	0.52	0.28	0.28	0.28	0.35	0.47	0.53	0.50	0.55
	13.5	0.38	0.38	0.60	0.76	0.72	0.57	0.82	0.43	0.49	0.34	0.42	0.51	0.57	0.32	0.31	0.32	0.38	0.50	0.57	0.54	0.59
	14.0	0.42	0.42	0.64	0.81	0.76	0.61	0.86	0.46	0.53	0.38	0.46	0.55	0.61	0.36	0.34	0.35	0.41	0.54	0.60	0.57	0.62
	14.5	0.45	0.45	0.68	0.85	0.80	0.65	0.91	0.49	0.57	0.42	0.50	0.59	0.65	0.40	0.38	0.38	0.45	0.57	0.64	0.61	0.66
	15.0	0.49	0.49	0.72	0.89	0.85	0.70	0.95	0.53	0.61	0.46	0.54	0.63	0.69	0.44	0.41	0.41	0.48	0.61	0.67	0.64	0.69

Table 1 Effect sizes for PAT scores — *continued*

		PAT Mathematics							PAT Reading Comprehension							PAT Reading Vocabulary						
Current year level		4	5	6	7	8	9	10	4	5	6	7	8	9	10	4	5	6	7	8	9	10
Mean PAT score		31.8	40.3	46.4	50.9	55.5	62.6	66.6	28.8	35.8	45.0	53.2	60.4	67.0	76.5	32.4	40.9	48.7	55.0	60.1	65.7	70.5
SD of PAT score		13.1	13.2	12.4	11.8	12.3	11.3	11.6	15.2	13.2	12.7	12.6	12.3	12.2	12.4	16.0	15.8	15.0	14.3	14.8	14.7	14.8
Difference between mean scores	15.5	0.53	0.53	0.76	0.93	0.89	0.74	0.99	0.56	0.64	0.50	0.58	0.67	0.73	0.48	0.44	0.44	0.51	0.64	0.70	0.67	0.72
	16.0	0.57	0.57	0.80	0.97	0.93	0.79	1.03	0.59	0.68	0.54	0.62	0.72	0.77	0.52	0.47	0.47	0.55	0.68	0.74	0.71	0.76
	16.5	0.61	0.61	0.84	1.02	0.97	0.83	1.08	0.63	0.72	0.57	0.66	0.76	0.81	0.56	0.50	0.51	0.58	0.71	0.77	0.74	0.79
	17.0	0.64	0.64	0.88	1.06	1.01	0.88	1.12	0.66	0.76	0.61	0.70	0.80	0.85	0.60	0.53	0.54	0.61	0.75	0.80	0.78	0.82
	17.5	0.68	0.68	0.92	1.10	1.05	0.92	1.16	0.69	0.80	0.65	0.74	0.84	0.89	0.65	0.56	0.57	0.65	0.78	0.84	0.81	0.86
	18.0	0.72	0.72	0.96	1.14	1.09	0.96	1.21	0.72	0.83	0.69	0.78	0.88	0.93	0.69	0.59	0.60	0.68	0.82	0.87	0.84	0.89
	18.5	0.76	0.76	1.00	1.19	1.13	1.01	1.25	0.76	0.87	0.73	0.82	0.92	0.98	0.73	0.63	0.63	0.71	0.85	0.91	0.88	0.93
	19.0	0.80	0.80	1.04	1.23	1.17	1.05	1.29	0.79	0.91	0.77	0.86	0.96	1.02	0.77	0.66	0.66	0.75	0.89	0.94	0.91	0.96
	19.5	0.83	0.83	1.08	1.27	1.21	1.10	1.34	0.82	0.95	0.81	0.90	1.00	1.06	0.81	0.69	0.70	0.78	0.92	0.97	0.95	0.99
	20.0	0.87	0.87	1.12	1.31	1.25	1.14	1.38	0.86	0.98	0.85	0.94	1.04	1.10	0.85	0.72	0.73	0.81	0.96	1.01	0.98	1.03
	20.5	0.91	0.91	1.16	1.36	1.29	1.19	1.42	0.89	1.02	0.89	0.98	1.08	1.14	0.89	0.75	0.76	0.85	0.99	1.04	1.01	1.06
	21.0	0.95	0.95	1.20	1.40	1.33	1.23	1.47	0.92	1.06	0.93	1.02	1.12	1.18	0.93	0.78	0.79	0.88	1.03	1.07	1.05	1.09
	21.5	0.98	0.98	1.24	1.44	1.37	1.27	1.51	0.95	1.10	0.97	1.06	1.16	1.22	0.97	0.81	0.82	0.91	1.06	1.11	1.08	1.13
	22.0	1.02	1.02	1.28	1.48	1.41	1.32	1.55	0.99	1.14	1.01	1.10	1.20	1.26	1.01	0.84	0.85	0.95	1.10	1.14	1.12	1.16
	22.5	1.06	1.06	1.32	1.53	1.46	1.36	1.59	1.02	1.17	1.05	1.13	1.24	1.30	1.05	0.88	0.89	0.98	1.13	1.18	1.15	1.20
	23.0	1.10	1.10	1.36	1.57	1.50	1.41	1.64	1.05	1.21	1.09	1.17	1.28	1.34	1.09	0.91	0.92	1.01	1.17	1.21	1.18	1.23
	23.5	1.14	1.14	1.40	1.61	1.54	1.45	1.68	1.09	1.25	1.13	1.21	1.33	1.39	1.13	0.94	0.95	1.05	1.20	1.24	1.22	1.26
	24.0	1.17	1.17	1.44	1.65	1.58	1.50	1.72	1.12	1.29	1.17	1.25	1.37	1.43	1.17	0.97	0.98	1.08	1.24	1.28	1.25	1.30
	24.5	1.21	1.21	1.48	1.69	1.62	1.54	1.77	1.15	1.33	1.20	1.29	1.41	1.47	1.21	1.00	1.01	1.11	1.27	1.31	1.29	1.33
	25.0	1.25	1.25	1.52	1.74	1.66	1.58	1.81	1.18	1.36	1.24	1.33	1.45	1.51	1.25	1.03	1.04	1.15	1.31	1.34	1.32	1.36

Table 1 Effect sizes for PAT scores — *continued*

		PAT Mathematics							PAT Reading Comprehension							PAT Reading Vocabulary						
Current year level		4	5	6	7	8	9	10	4	5	6	7	8	9	10	4	5	6	7	8	9	10
Mean PAT score		31.8	40.3	46.4	50.9	55.5	62.6	66.6	28.8	35.8	45.0	53.2	60.4	67.0	76.5	32.4	40.9	48.7	55.0	60.1	65.7	70.5
SD of PAT score		13.1	13.2	12.4	11.8	12.3	11.3	11.6	15.2	13.2	12.7	12.6	12.3	12.2	12.4	16.0	15.8	15.0	14.3	14.8	14.7	14.8
Difference between mean scores	25.5	1.29	1.29	1.56	1.78	1.70	1.63	1.85	1.22	1.40	1.28	1.37	1.49	1.55	1.29	1.06	1.08	1.18	1.34	1.38	1.35	1.40
	26.0	1.33	1.33	1.60	1.82	1.74	1.67	1.90	1.25	1.44	1.32	1.41	1.53	1.59	1.33	1.09	1.11	1.21	1.38	1.41	1.39	1.43
	26.5	1.36	1.36	1.65	1.86	1.78	1.72	1.94	1.28	1.48	1.36	1.45	1.57	1.63	1.37	1.13	1.14	1.25	1.41	1.45	1.42	1.47
	27.0	1.40	1.40	1.69	1.91	1.82	1.76	1.98	1.32	1.52	1.40	1.49	1.61	1.67	1.41	1.16	1.17	1.28	1.45	1.48	1.46	1.50
	27.5	1.44	1.44	1.73	1.95	1.86	1.81	2.03	1.35	1.55	1.44	1.53	1.65	1.71	1.45	1.19	1.20	1.31	1.48	1.51	1.49	1.53
	28.0	1.48	1.48	1.77	1.99	1.90	1.85	2.07	1.38	1.59	1.48	1.57	1.69	1.75	1.49	1.22	1.23	1.35	1.52	1.55	1.52	1.57
	28.5	1.52	1.52	1.81	2.03	1.94	1.89	2.11	1.41	1.63	1.52	1.61	1.73	1.80	1.53	1.25	1.27	1.38	1.55	1.58	1.56	1.60
	29.0	1.55	1.55	1.85	2.08	1.98	1.94	2.16	1.45	1.67	1.56	1.65	1.77	1.84	1.57	1.28	1.30	1.41	1.59	1.61	1.59	1.64
	29.5	1.59	1.59	1.89	2.12	2.02	1.98	2.20	1.48	1.70	1.60	1.69	1.81	1.88	1.61	1.31	1.33	1.45	1.62	1.65	1.63	1.67
	30.0	1.63	1.63	1.93	2.16	2.07	2.03	2.24	1.51	1.74	1.64	1.73	1.85	1.92	1.65	1.34	1.36	1.48	1.66	1.68	1.66	1.70
	30.5	1.67	1.67	1.97	2.20	2.11	2.07	2.28	1.55	1.78	1.68	1.77	1.89	1.96	1.69	1.38	1.39	1.51	1.69	1.72	1.69	1.74
	31.0	1.70	1.70	2.01	2.25	2.15	2.12	2.33	1.58	1.82	1.72	1.81	1.93	2.00	1.73	1.41	1.42	1.55	1.73	1.75	1.73	1.77
	31.5	1.74	1.74	2.05	2.29	2.19	2.16	2.37	1.61	1.86	1.76	1.85	1.98	2.04	1.77	1.44	1.46	1.58	1.76	1.78	1.76	1.80
	32.0	1.78	1.78	2.09	2.33	2.23	2.20	2.41	1.64	1.89	1.80	1.89	2.02	2.08	1.81	1.47	1.49	1.61	1.80	1.82	1.80	1.84
	32.5	1.82	1.82	2.13	2.37	2.27	2.25	2.46	1.68	1.93	1.83	1.93	2.06	2.12	1.85	1.50	1.52	1.65	1.83	1.85	1.83	1.87
	33.0	1.86	1.86	2.17	2.42	2.31	2.29	2.50	1.71	1.97	1.87	1.97	2.10	2.16	1.90	1.53	1.55	1.68	1.87	1.89	1.86	1.91

Table 2 Effect sizes for STAR test scores

STAR scores are stanines, so the expected change is 0, and standard deviation is 2.

Stanine difference	Effect size
0.2	0.1
0.4	0.2
0.6	0.3
0.8	0.4
1	0.5
1.2	0.6
1.4	0.7
1.6	0.8
1.8	0.9
2	1
2.2	1.1
2.4	1.2
2.6	1.3
2.8	1.4
3	1.5
3.2	1.6
3.4	1.7
3.6	1.8
3.8	1.9
4	2
4.2	2.1
4.4	2.2
4.6	2.3
4.8	2.4
5	2.5
5.2	2.6

Table 3 Effect sizes for asTTle Mathematics scores

The effect size values in the table are for differences across **one year**. The calculations are based on a standard deviation of 100.

		asTTle Mathematics								
Current year level		4	5	6	7	8	9	10	11	12
Mean asTTle score		410	470	502	541	638	774	807	828	848
Difference between mean scores	44.0	-0.16	-0.16	0.12	0.05	-0.53	-0.92	0.11	0.23	0.24
	46.0	-0.14	-0.14	0.14	0.07	-0.51	-0.90	0.13	0.25	0.26
	48.0	-0.12	-0.12	0.16	0.09	-0.49	-0.88	0.15	0.27	0.28
	50.0	-0.10	-0.10	0.18	0.11	-0.47	-0.86	0.17	0.29	0.30
	52.0	-0.08	-0.08	0.20	0.13	-0.45	-0.84	0.19	0.31	0.32
	54.0	-0.06	-0.06	0.22	0.15	-0.43	-0.82	0.21	0.33	0.34
	56.0	-0.04	-0.04	0.24	0.17	-0.41	-0.80	0.23	0.35	0.36
	58.0	-0.02	-0.02	0.26	0.19	-0.39	-0.78	0.25	0.37	0.38
	60.0	0.00	0.00	0.28	0.21	-0.37	-0.76	0.27	0.39	0.40
	62.0	0.02	0.02	0.30	0.23	-0.35	-0.74	0.29	0.41	0.42
	64.0	0.04	0.04	0.32	0.25	-0.33	-0.72	0.31	0.43	0.44
	66.0	0.06	0.06	0.34	0.27	-0.31	-0.70	0.33	0.45	0.46
	68.0	0.08	0.08	0.36	0.29	-0.29	-0.68	0.35	0.47	0.48
	70.0	0.10	0.10	0.38	0.31	-0.27	-0.66	0.37	0.49	0.50
	72.0	0.12	0.12	0.40	0.33	-0.25	-0.64	0.39	0.51	0.52
	74.0	0.14	0.14	0.42	0.35	-0.23	-0.62	0.41	0.53	0.54
	76.0	0.16	0.16	0.44	0.37	-0.21	-0.60	0.43	0.55	0.56
	78.0	0.18	0.18	0.46	0.39	-0.19	-0.58	0.45	0.57	0.58
	80.0	0.20	0.20	0.48	0.41	-0.17	-0.56	0.47	0.59	0.60
	82.0	0.22	0.22	0.50	0.43	-0.15	-0.54	0.49	0.61	0.62
	84.0	0.24	0.24	0.52	0.45	-0.13	-0.52	0.51	0.63	0.64
	86.0	0.26	0.26	0.54	0.47	-0.11	-0.50	0.53	0.65	0.66
	88.0	0.28	0.28	0.56	0.49	-0.09	-0.48	0.55	0.67	0.68
	90.0	0.30	0.30	0.58	0.51	-0.07	-0.46	0.57	0.69	0.70
	92.0	0.32	0.32	0.60	0.53	-0.05	-0.44	0.59	0.71	0.72
	94.0	0.34	0.34	0.62	0.55	-0.03	-0.42	0.61	0.73	0.74
	96.0	0.36	0.36	0.64	0.57	-0.01	-0.40	0.63	0.75	0.76
	98.0	0.38	0.38	0.66	0.59	0.01	-0.38	0.65	0.77	0.78
	100.0	0.40	0.40	0.68	0.61	0.03	-0.36	0.67	0.79	0.80
	105.0	0.45	0.45	0.73	0.66	0.08	-0.31	0.72	0.84	0.85
	110.0	0.50	0.50	0.78	0.71	0.13	-0.26	0.77	0.89	0.90
	115.0	0.55	0.55	0.83	0.76	0.18	-0.21	0.82	0.94	0.95
	120.0	0.60	0.60	0.88	0.81	0.23	-0.16	0.87	0.99	1.00
	125.0	0.65	0.65	0.93	0.86	0.28	-0.11	0.92	1.04	1.05
	130.0	0.70	0.70	0.98	0.91	0.33	-0.06	0.97	1.09	1.10
	135.0	0.75	0.75	1.03	0.96	0.38	-0.01	1.02	1.14	1.15

Table 3 Effect sizes for asTTle Mathematics scores — *continued*

		asTTle Mathematics								
Current year level		4	5	6	7	8	9	10	11	12
Mean asTTle score		410	470	502	541	638	774	807	828	848
Difference between mean scores	140.0	0.80	0.80	1.08	1.01	0.43	0.04	1.07	1.19	1.20
	145.0	0.85	0.85	1.13	1.06	0.48	0.09	1.12	1.24	1.25
	150.0	0.90	0.90	1.18	1.11	0.53	0.14	1.17	1.29	1.30
	155.0	0.95	0.95	1.23	1.16	0.58	0.19	1.22	1.34	1.35
	160.0	1.00	1.00	1.28	1.21	0.63	0.24	1.27	1.39	1.40
	165.0	1.05	1.05	1.33	1.26	0.68	0.29	1.32	1.44	1.45
	170.0	1.10	1.10	1.38	1.31	0.73	0.34	1.37	1.49	1.50
	175.0	1.15	1.15	1.43	1.36	0.78	0.39	1.42	1.54	1.55
	180.0	1.20	1.20	1.48	1.41	0.83	0.44	1.47	1.59	1.60
	185.0	1.25	1.25	1.53	1.46	0.88	0.49	1.52	1.64	1.65
	190.0	1.30	1.30	1.58	1.51	0.93	0.54	1.57	1.69	1.70
	195.0	1.35	1.35	1.63	1.56	0.98	0.59	1.62	1.74	1.75
	200.0	1.40	1.40	1.68	1.61	1.03	0.64	1.67	1.79	1.80
	210.0	1.50	1.50	1.78	1.71	1.13	0.74	1.77	1.89	1.90
	220.0	1.60	1.60	1.88	1.81	1.23	0.84	1.87	1.99	2.00
	230.0	1.70	1.70	1.98	1.91	1.33	0.94	1.97	2.09	2.10
	240.0	1.80	1.80	2.08	2.01	1.43	1.04	2.07	2.19	2.20

Table 4 Effect sizes for asTTle Reading scores

The effect size values in the table are for differences across **one year**. The calculations are based on a standard deviation of 100.

		asTTle Reading								
Current year level		4	5	6	7	8	9	10	11	12
Mean asTTle score		412	462	489	508	517	634	728	768	780
Difference between mean scores	44.0	-0.06	-0.06	0.17	0.25	0.35	-0.73	-0.50	0.04	0.32
	46.0	-0.04	-0.04	0.19	0.27	0.37	-0.71	-0.48	0.06	0.34
	48.0	-0.02	-0.02	0.21	0.29	0.39	-0.69	-0.46	0.08	0.36
	50.0	0.00	0.00	0.23	0.31	0.41	-0.67	-0.44	0.10	0.38
	52.0	0.02	0.02	0.25	0.33	0.43	-0.65	-0.42	0.12	0.40
	54.0	0.04	0.04	0.27	0.35	0.45	-0.63	-0.40	0.14	0.42
	56.0	0.06	0.06	0.29	0.37	0.47	-0.61	-0.38	0.16	0.44
	58.0	0.08	0.08	0.31	0.39	0.49	-0.59	-0.36	0.18	0.46
	60.0	0.10	0.10	0.33	0.41	0.51	-0.57	-0.34	0.20	0.48
	62.0	0.12	0.12	0.35	0.43	0.53	-0.55	-0.32	0.22	0.50
	64.0	0.14	0.14	0.37	0.45	0.55	-0.53	-0.30	0.24	0.52
	66.0	0.16	0.16	0.39	0.47	0.57	-0.51	-0.28	0.26	0.54
	68.0	0.18	0.18	0.41	0.49	0.59	-0.49	-0.26	0.28	0.56
	70.0	0.20	0.20	0.43	0.51	0.61	-0.47	-0.24	0.30	0.58
	72.0	0.22	0.22	0.45	0.53	0.63	-0.45	-0.22	0.32	0.60
	74.0	0.24	0.24	0.47	0.55	0.65	-0.43	-0.20	0.34	0.62
	76.0	0.26	0.26	0.49	0.57	0.67	-0.41	-0.18	0.36	0.64
	78.0	0.28	0.28	0.51	0.59	0.69	-0.39	-0.16	0.38	0.66
	80.0	0.30	0.30	0.53	0.61	0.71	-0.37	-0.14	0.40	0.68
	82.0	0.32	0.32	0.55	0.63	0.73	-0.35	-0.12	0.42	0.70
	84.0	0.34	0.34	0.57	0.65	0.75	-0.33	-0.10	0.44	0.72
	86.0	0.36	0.36	0.59	0.67	0.77	-0.31	-0.08	0.46	0.74
	88.0	0.38	0.38	0.61	0.69	0.79	-0.29	-0.06	0.48	0.76
	90.0	0.40	0.40	0.63	0.71	0.81	-0.27	-0.04	0.50	0.78
	92.0	0.42	0.42	0.65	0.73	0.83	-0.25	-0.02	0.52	0.80
	94.0	0.44	0.44	0.67	0.75	0.85	-0.23	0.00	0.54	0.82
	96.0	0.46	0.46	0.69	0.77	0.87	-0.21	0.02	0.56	0.84
	98.0	0.48	0.48	0.71	0.79	0.89	-0.19	0.04	0.58	0.86
	100.0	0.50	0.50	0.73	0.81	0.91	-0.17	0.06	0.60	0.88
	105.0	0.55	0.55	0.78	0.86	0.96	-0.12	0.11	0.65	0.93
	110.0	0.60	0.60	0.83	0.91	1.01	-0.07	0.16	0.70	0.98
	115.0	0.65	0.65	0.88	0.96	1.06	-0.02	0.21	0.75	1.03
	120.0	0.70	0.70	0.93	1.01	1.11	0.03	0.26	0.80	1.08
	125.0	0.75	0.75	0.98	1.06	1.16	0.08	0.31	0.85	1.13
	130.0	0.80	0.80	1.03	1.11	1.21	0.13	0.36	0.90	1.18
	135.0	0.85	0.85	1.08	1.16	1.26	0.18	0.41	0.95	1.23

Table 4 Effect sizes for asTTle Reading scores — *continued*

		asTTle Reading								
Current year level		4	5	6	7	8	9	10	11	12
Mean asTTle score		412	462	489	508	517	634	728	768	780
Difference between mean scores	140.0	0.90	0.90	1.13	1.21	1.31	0.23	0.46	1.00	1.28
	145.0	0.95	0.95	1.18	1.26	1.36	0.28	0.51	1.05	1.33
	150.0	1.00	1.00	1.23	1.31	1.41	0.33	0.56	1.10	1.38
	155.0	1.05	1.05	1.28	1.36	1.46	0.38	0.61	1.15	1.43
	160.0	1.10	1.10	1.33	1.41	1.51	0.43	0.66	1.20	1.48
	165.0	1.15	1.15	1.38	1.46	1.56	0.48	0.71	1.25	1.53
	170.0	1.20	1.20	1.43	1.51	1.61	0.53	0.76	1.30	1.58
	175.0	1.25	1.25	1.48	1.56	1.66	0.58	0.81	1.35	1.63
	180.0	1.30	1.30	1.53	1.61	1.71	0.63	0.86	1.40	1.68
	185.0	1.35	1.35	1.58	1.66	1.76	0.68	0.91	1.45	1.73
	190.0	1.40	1.40	1.63	1.71	1.81	0.73	0.96	1.50	1.78
	195.0	1.45	1.45	1.68	1.76	1.86	0.78	1.01	1.55	1.83
	200.0	1.50	1.50	1.73	1.81	1.91	0.83	1.06	1.60	1.88
	210.0	1.60	1.60	1.83	1.91	2.01	0.93	1.16	1.70	1.98
	220.0	1.70	1.70	1.93	2.01	2.11	1.03	1.26	1.80	2.08
	230.0	1.80	1.80	2.03	2.11	2.21	1.13	1.36	1.90	2.18
	240.0	1.90	1.90	2.13	2.21	2.31	1.23	1.46	2.00	2.28

Table 5 Effect sizes for asTTle Writing scores

The effect size values in the table are for differences across **one year**. The calculations are based on a standard deviation of 100.

		asTTle Writing								
Current year level		4	5	6	7	8	9	10	11	12
Mean asTTle score		454	482	504	518	536	590	633	639	669
Difference between mean scores	44.0	0.16	0.16	0.22	0.30	0.26	-0.10	0.01	0.38	0.14
	46.0	0.18	0.18	0.24	0.32	0.28	-0.08	0.03	0.40	0.16
	48.0	0.20	0.20	0.26	0.34	0.30	-0.06	0.05	0.42	0.18
	50.0	0.22	0.22	0.28	0.36	0.32	-0.04	0.07	0.44	0.20
	52.0	0.24	0.24	0.30	0.38	0.34	-0.02	0.09	0.46	0.22
	54.0	0.26	0.26	0.32	0.40	0.36	0.00	0.11	0.48	0.24
	56.0	0.28	0.28	0.34	0.42	0.38	0.02	0.13	0.50	0.26
	58.0	0.30	0.30	0.36	0.44	0.40	0.04	0.15	0.52	0.28
	60.0	0.32	0.32	0.38	0.46	0.42	0.06	0.17	0.54	0.30
	62.0	0.34	0.34	0.40	0.48	0.44	0.08	0.19	0.56	0.32
	64.0	0.36	0.36	0.42	0.50	0.46	0.10	0.21	0.58	0.34
	66.0	0.38	0.38	0.44	0.52	0.48	0.12	0.23	0.60	0.36
	68.0	0.40	0.40	0.46	0.54	0.50	0.14	0.25	0.62	0.38
	70.0	0.42	0.42	0.48	0.56	0.52	0.16	0.27	0.64	0.40
	72.0	0.44	0.44	0.50	0.58	0.54	0.18	0.29	0.66	0.42
	74.0	0.46	0.46	0.52	0.60	0.56	0.20	0.31	0.68	0.44
	76.0	0.48	0.48	0.54	0.62	0.58	0.22	0.33	0.70	0.46
	78.0	0.50	0.50	0.56	0.64	0.60	0.24	0.35	0.72	0.48
	80.0	0.52	0.52	0.58	0.66	0.62	0.26	0.37	0.74	0.50
	82.0	0.54	0.54	0.60	0.68	0.64	0.28	0.39	0.76	0.52
	84.0	0.56	0.56	0.62	0.70	0.66	0.30	0.41	0.78	0.54
	86.0	0.58	0.58	0.64	0.72	0.68	0.32	0.43	0.80	0.56
	88.0	0.60	0.60	0.66	0.74	0.70	0.34	0.45	0.82	0.58
	90.0	0.62	0.62	0.68	0.76	0.72	0.36	0.47	0.84	0.60
	92.0	0.64	0.64	0.70	0.78	0.74	0.38	0.49	0.86	0.62
	94.0	0.66	0.66	0.72	0.80	0.76	0.40	0.51	0.88	0.64
	96.0	0.68	0.68	0.74	0.82	0.78	0.42	0.53	0.90	0.66
	98.0	0.70	0.70	0.76	0.84	0.80	0.44	0.55	0.92	0.68
	100.0	0.72	0.72	0.78	0.86	0.82	0.46	0.57	0.94	0.70
	105.0	0.77	0.77	0.83	0.91	0.87	0.51	0.62	0.99	0.75
	110.0	0.82	0.82	0.88	0.96	0.92	0.56	0.67	1.04	0.80
	115.0	0.87	0.87	0.93	1.01	0.97	0.61	0.72	1.09	0.85
	120.0	0.92	0.92	0.98	1.06	1.02	0.66	0.77	1.14	0.90
	125.0	0.97	0.97	1.03	1.11	1.07	0.71	0.82	1.19	0.95
	130.0	1.02	1.02	1.08	1.16	1.12	0.76	0.87	1.24	1.00
	135.0	1.07	1.07	1.13	1.21	1.17	0.81	0.92	1.29	1.05

Table 5 Effect sizes for asTTle writing scores — *continued*

		asTTle writing								
Current Year level		4	5	6	7	8	9	10	11	12
Mean asTTle score		454	482	504	518	536	590	633	639	669
Difference between mean scores	140.0	1.12	1.12	1.18	1.26	1.22	0.86	0.97	1.34	1.10
	145.0	1.17	1.17	1.23	1.31	1.27	0.91	1.02	1.39	1.15
	150.0	1.22	1.22	1.28	1.36	1.32	0.96	1.07	1.44	1.20
	155.0	1.27	1.27	1.33	1.41	1.37	1.01	1.12	1.49	1.25
	160.0	1.32	1.32	1.38	1.46	1.42	1.06	1.17	1.54	1.30
	165.0	1.37	1.37	1.43	1.51	1.47	1.11	1.22	1.59	1.35
	170.0	1.42	1.42	1.48	1.56	1.52	1.16	1.27	1.64	1.40
	175.0	1.47	1.47	1.53	1.61	1.57	1.21	1.32	1.69	1.45
	180.0	1.52	1.52	1.58	1.66	1.62	1.26	1.37	1.74	1.50
	185.0	1.57	1.57	1.63	1.71	1.67	1.31	1.42	1.79	1.55
	190.0	1.62	1.62	1.68	1.76	1.72	1.36	1.47	1.84	1.60
	195.0	1.67	1.67	1.73	1.81	1.77	1.41	1.52	1.89	1.65
	200.0	1.72	1.72	1.78	1.86	1.82	1.46	1.57	1.94	1.70
	210.0	1.82	1.82	1.88	1.96	1.92	1.56	1.67	2.04	1.80
	220.0	1.92	1.92	1.98	2.06	2.02	1.66	1.77	2.14	1.90
	230.0	2.02	2.02	2.08	2.16	2.12	1.76	1.87	2.24	2.00
	240.0	2.12	2.12	2.18	2.26	2.22	1.86	1.97	2.34	2.10

Table 6 Standard error of effect size estimates: two independent samples (different students in the two tests)

n_2	n_1																	
	10	15	20	25	30	35	40	50	60	75	100	150	200	250	300	500	1000	2000
10	0.45	0.41	0.39	0.37	0.37	0.36	0.35	0.35	0.34	0.34	0.33	0.33	0.32	0.32	0.32	0.32	0.32	0.32
15	0.41	0.37	0.34	0.33	0.32	0.31	0.30	0.29	0.29	0.28	0.28	0.27	0.27	0.27	0.26	0.26	0.26	0.26
20	0.39	0.34	0.32	0.30	0.29	0.28	0.27	0.26	0.26	0.25	0.24	0.24	0.23	0.23	0.23	0.23	0.23	0.22
25	0.37	0.33	0.30	0.28	0.27	0.26	0.25	0.24	0.24	0.23	0.22	0.22	0.21	0.21	0.21	0.20	0.20	0.20
30	0.37	0.32	0.29	0.27	0.26	0.25	0.24	0.23	0.22	0.22	0.21	0.20	0.20	0.19	0.19	0.19	0.19	0.18
35	0.36	0.31	0.28	0.26	0.25	0.24	0.23	0.22	0.21	0.20	0.20	0.19	0.18	0.18	0.18	0.17	0.17	0.17
40	0.35	0.30	0.27	0.25	0.24	0.23	0.22	0.21	0.20	0.20	0.19	0.18	0.17	0.17	0.17	0.16	0.16	0.16
50	0.35	0.29	0.26	0.24	0.23	0.22	0.21	0.20	0.19	0.18	0.17	0.16	0.16	0.15	0.15	0.15	0.14	0.14
60	0.34	0.29	0.26	0.24	0.22	0.21	0.20	0.19	0.18	0.17	0.16	0.15	0.15	0.14	0.14	0.14	0.13	0.13
75	0.34	0.28	0.25	0.23	0.22	0.20	0.20	0.18	0.17	0.16	0.15	0.14	0.14	0.13	0.13	0.12	0.12	0.12
100	0.33	0.28	0.24	0.22	0.21	0.20	0.19	0.17	0.16	0.15	0.14	0.13	0.12	0.12	0.12	0.11	0.10	0.10
150	0.33	0.27	0.24	0.22	0.20	0.19	0.18	0.16	0.15	0.14	0.13	0.12	0.11	0.10	0.10	0.09	0.09	0.08
200	0.32	0.27	0.23	0.21	0.20	0.18	0.17	0.16	0.15	0.14	0.12	0.11	0.10	0.09	0.09	0.08	0.08	0.07
250	0.32	0.27	0.23	0.21	0.19	0.18	0.17	0.15	0.14	0.13	0.12	0.10	0.09	0.09	0.09	0.08	0.07	0.07
300	0.32	0.26	0.23	0.21	0.19	0.18	0.17	0.15	0.14	0.13	0.12	0.10	0.09	0.09	0.08	0.07	0.07	0.06
500	0.32	0.26	0.23	0.20	0.19	0.17	0.16	0.15	0.14	0.12	0.11	0.09	0.08	0.08	0.07	0.06	0.05	0.05
1000	0.32	0.26	0.23	0.20	0.19	0.17	0.16	0.14	0.13	0.12	0.10	0.09	0.08	0.07	0.07	0.05	0.04	0.04
2000	0.32	0.26	0.22	0.20	0.18	0.17	0.16	0.14	0.13	0.12	0.10	0.08	0.07	0.07	0.06	0.05	0.04	0.03

Note that it does not matter which sample is regarded as sample 1 and which sample 2; the table is symmetric (the values in the first row equal those in the first column, etc.).

Table 7 Standard error of effect size estimates: two dependent samples (matched students in the two tests)

STAR raw scores (and stanine scores) tend to have correlations (*r*) of between 0.8 and 0.9; PAT scores tend to have correlations of between 0.7 (between Year 3 and Year 4 scores) and 0.9 (all other years); and asTTle scores tend to have correlations between 0.6 and 0.9 (most correlations are in the 0.6 to 0.7 range). Use this information to pick the most appropriate column/s for your data.

	r								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	0.42	0.40	0.37	0.35	0.32	0.28	0.24	0.20	0.14
15	0.35	0.33	0.31	0.28	0.26	0.23	0.20	0.16	0.12
20	0.30	0.28	0.26	0.24	0.22	0.20	0.17	0.14	0.10
25	0.27	0.25	0.24	0.22	0.20	0.18	0.15	0.13	0.09
30	0.24	0.23	0.22	0.20	0.18	0.16	0.14	0.12	0.08
35	0.23	0.21	0.20	0.19	0.17	0.15	0.13	0.11	0.08
40	0.21	0.20	0.19	0.17	0.16	0.14	0.12	0.10	0.07
50	0.19	0.18	0.17	0.15	0.14	0.13	0.11	0.09	0.06
60	0.17	0.16	0.15	0.14	0.13	0.12	0.10	0.08	0.06
75	0.15	0.15	0.14	0.13	0.12	0.10	0.09	0.07	0.05
100	0.13	0.13	0.12	0.11	0.10	0.09	0.08	0.06	0.04
150	0.11	0.10	0.10	0.09	0.08	0.07	0.06	0.05	0.04
200	0.09	0.09	0.08	0.08	0.07	0.06	0.05	0.04	0.03
250	0.08	0.08	0.07	0.07	0.06	0.06	0.05	0.04	0.03
300	0.08	0.07	0.07	0.06	0.06	0.05	0.04	0.04	0.03
500	0.06	0.06	0.05	0.05	0.04	0.04	0.03	0.03	0.02
1000	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.01
2000	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.01

Acknowledgements

Thanks are due to John Hattie of Auckland University, and colleagues in the Research Division at the Ministry of Education and at NZCER who have commented on the content of this paper at various times. However, ultimately the opinions expressed in this paper belong to the authors alone and are not to be construed as official policy from either organisation.

Ian Schagen
Edith Hodgen
March 2009

Dr Ian Schagen was Head of Statistics at the National Foundation for Educational Research in the UK until April 2008, when he came to New Zealand for a year, working with the Research Division of the Ministry of Education as a Chief Research Analyst.

Edith Hodgen is the manager of the Statistics and Data Management team at NZCER.