Understanding, using and calculating effect size

"Effect size … allows us to move beyond the simplistic, "Does it work or not?" to the far more sophisticated, "How well does it work in a range of contexts?" (Coe, 2002)

What is effect size?

Effect size is a simple measure for quantifying the difference between two groups or the same group over time, on a common scale.

In an educational setting, effect size is one way to measure the effectiveness of a particular intervention. Effect size enables us to measure both the *improvement* (gain) in learner achievement for a group of learners AND the *variation* of student performances expressed on a standardised scale. By taking into account both *improvement* and *variation* it provides information about which interventions are worth having.

Influence	Effect Size
Feedback	0.75
Teacher- student relationships	0.72
Meta-cognitive strategies	0.69
Direct instruction	0.59
Homework	0.29
Reducing class size	0.21
Ability grouping/ tracking	0.12

Dr John Hattie, in his analysis of hundreds of international and national educational interventions and data, determined that *"for students moving from one year to the next, the average effect size across all students is 0.40."* Hattie's research places particular emphasis on programs with effect sizes above 0.4 as worth having and those lower than 0.4 as needing further consideration (refer to the table indicated and Appendix 2). It should be noted that the 0.4 "hinge point" used by Hattie is an average of many measures and so should be used as a guide only. It is more appropriate to compare local school effect sizes with the corresponding equivalent group or state level effect size.

How is effect size calculated?

Effect size is calculated by taking the difference in two mean scores and then dividing this figure by the average spread of student scores (i.e. average standard deviation*). To be valid, the spread of scores should be approximately distributed in a 'normal' bell curve shape. See formula below.

$$\text{Effect Size (ES)} = \frac{\text{Average of the post-test scores} - \text{Average of the pre-test scores}}{\text{Average standard deviation}^*}$$

**The average standard deviation in the above formula refers to the standard deviation for the pre-test and post-test data calculated individually, then averaged. A complete example using MS Excel to do the calculation is provided in Appendix 1.*

How can we use effect size?

There are many ways in which to use effect sizes. This resource focuses on using and understanding effect sizes to:

- Investigate the effectiveness of a particular intervention for a defined group of students
- Compare the effectiveness of different interventions
- Evaluate the growth over time.

Example: A curriculum leader is using effect size to understand and estimate the impact of a particular approach to reading comprehension by comparing achievement scores using PAT R Comprehension (or equivalent assessment) for the same students over a year. In reviewing the school's PAT R effect size results for the same students from *Year 5, Term 3, 2010* to *Year 6, Term 3, 2011* an effect size of 0.49 is recorded, but effect sizes for individual classes are 0.86, 0.42 and 0.18 respectively. This indicates that more than the expected average progress is being made, and raises questions listed below, aimed at achieving greater effectiveness and consistency.

What questions can we ask?

The most important consideration when using effect size are the questions it raises. It invites educators to reflect on:

- *"How well is what I am doing working for different groups of students each year and why?"*
- *"What possible reasons could there be for some student or groups of students progressing more or less?"*
- *"How does student progress compare with their achievement levels"?*

These questions lead to more focussed investigation about the effectiveness of what we do. This provides a basis for teaching and learning interventions we should *stop*, *start* or *continue* as part of effective educational practice.

Draft

How can effect size be used reliably?

Multiple measures are still required

> *"Comparing results on different measures gives teachers insights into what teaching strategies, as well as testing strategies, work best with different students." (Bernhardt, 2004)*

Effect size is only a single measure of progress and DIA*f* self review processes encourage educators to *use a range of learner achievement and multiple measures of data* to complement existing achievement measures in order to reliably understand and replicate evidence of what works. Bernhardt (2004) states that demographic, perception, student learning and process measures about the teaching and learning environment is what provides a comprehensive picture of what makes a difference to learners. It is difficult to draw any conclusions that a particular intervention is effective or ineffective using a single measure.

Caution for all small sample sizes and at the individual student level

Effect size for cohorts smaller than 30 are often not suitable for reliably estimating the impact of an intervention. Hattie suggests that care should be taken in the interpretation of any findings for small sample sizes as outliers in student scores can skew the effect sizes and may require special consideration. Effect sizes derived from small sample sizes and *individual student effect sizes* should only be used indicatively by the teacher to question - *What possible reasons could there be for why that group of students recorded these estimated effect sizes? What will we do for students who are achieving at expected achievement levels but not the expected growth effect size?* Interpretation of effect sizes for individual students is to be used with caution because we would expect larger errors in effect size at this level (refer to Appendix 1). Therefore individual level effects must always be used in addition with other reliable information and teacher professional judgement.

Accuracy is enhanced when comparing the exact same group of students

When comparing pre-test and post-test scores, it is most useful to ensure that all students are tested and that *scores from exactly the same group of students are compared.* Using students' ED ID ensures you are looking at the effect of an intervention on the same students who experienced the intervention over the period being considered. This enhances the accuracy and interpretation of the results.

NAPLAN effect sizes cannot be compared equally

NAPLAN effect sizes calculated for the Year 3-5 cohort should not be compared with Year 5-7 and Year 7-9 cohort effect sizes using the 0.4 average effect size interpretation. There are larger effect sizes for Year 3-5 than in Year 5-7 and Year 7-9. In addition, students at lower proficiency bands will tend to show greater gains than students in higher proficiency bands and care is needed for students that attain maximum or near maximum scores as it is difficult to show growth (due to this ceiling effect). It is recommended that NAPLAN effect size values only be compared over time for equivalent groups in the same school (e.g. Year 3-5), across statistically similar/like schools or with the corresponding state level effect size.

> *"Interpretation requires time, thoughtfulness, reservation of judgements and open challenge … it is formulating possibilities, developing convincing arguments, locating logical flaws and establishing a feasible and defensible notion of what the data represent" (Earl, 2006)*

In summary, it is important to base the interpretation of effect size on the full range of contextual and measurement factors. This measure is best used to raise questions in conversation and stimulate discussion, particularly around the possible reasons for differences and the question:

"What positive difference are we making for this group of students?"

References:

Bernhardt, V., (2004). Data Analysis for Continuous School Improvement. Eye on Education, Larchmont: NY.

Coe, R., (2002). *It's the Effect Size, Stupid. What effect size is and why it is important* presentation to the Annual Conference of the British Educational Research Association, England 2002. Retrieved November 2011 from http://www.leeds.ac.uk/educol/documents/00002182.htm

Earl, L. & Katz, S., (2006). Leading schools in a Data-Rich World: Harnessing data for school improvement. Corwin Press, California.

Hattie, J., (2012). Visible Learning for Teachers, Maximising Impact on Learning. Routledge, Oxford: UK.

Hattie, J., & Masters, D., (2011). *Visible Learning Plus. Supporting Material Visible Learning Workshop* presentation in Adelaide, South Australia, 2011.

Hattie, J., (2003). *"Teachers Make a Difference, What is the research evidence?". 2003 - Building Teacher Quality: What does the research tell us?* Retrieved November 2011 from http://research.acer.edu.au/research_conference_2003/4

Schagen, I., & Hodgen, E., (2009). *How Much Difference Does it make? Notes on Understanding, Using, and Calculating Effect Size for Schools.* Retrieved from *http://www.educationcounts.govt.nz/publications/schooling/36097/36098*

Draft

Appendix 1: Effect size calculation example

The following represents the 6 step process for calculating effect size manually and the corresponding formulas that can be used in MS Excel (indicated in shaded blue text) to calculate these statistical measures. The attached figure below contains individual student data for a typical assessment.

	A	B	C	D	E
2	Student ID	2010 score	2011 score	Diff	Effect Size
3	17	551	535	-16	-0.25
4	23	502	495	-7	-0.11
5	10	443	448	5	0.08
6	4	469	486	17	0.27
7	28	380	400	20	0.31
8	1	502	525	23	0.36
9	3	502	525	23	0.36
10	8	431	457	26	0.41
11	24	502	535	33	0.52
12	18	312	346	33	0.53
13	20	420	457	37	0.58
14	16	484	535	51	0.8
15	26	443	495	52	0.81
16	25	502	557	55	0.86
17	9	469	525	56	0.88
18	2	390	448	58	0.91
19	11	523	593	70	1.09
20	12	431	505	74	1.16
21	14	420	495	75	1.17
22	21	410	495	85	1.33
23	15	455	546	91	1.42
24	19	370	467	97	1.52
25	27	523	623	100	1.56
26	22	502	607	105	1.64
27	7	469	581	112	1.75
28	6	312	429	117	1.83
29	5	380	505	125	1.95
30	13	322	448	126	1.97
31	1) Average	**443.5** = AVERAGE(B3:B30)	**502.3** = AVERAGE(C3:C30)		
32	2) Difference			**58.7** = C31-B31	
33	3) Spread (Standard deviation or SD)	**65.7** = STDEV(B3:B30)	**62.2** = STDEV(C3:C30)		
34	4) Average sd			**64.0** =AVERAGE(B33:C33)	
35	5) Effect Size				**0.92** = D32/D34

1. Mean score (or average) is calculated by adding all the individual student scores together and then dividing by the total number of student scores. In the example provided (see attached Figure):

Mean score (for 2010) **= (551+502+443+ …+380+322) ÷ *28* =**
=AVERAGE(B3:B30) = 443.5 [See Excel Figure, cell B31]

Mean score (for 2011) **= (535+495+448+ …+505+448) ÷ *28* =**
=AVERAGE(C3:C30) = 502.3 [See Excel Figure, cell C31]

2. the difference between two mean scores (also referred to as the 'gain' score in ACARA NAPLAN resources)

=C31–B31* = 502.3 – 443.5 = 58.7 *[See Excel Figure, cell D32]

3. Standard Deviation (SD) can be a complicated formula to calculate manually (i.e. basically the average of the sum of the squared differences from the mean score) and can be easily calculated in MS Excel as follows:

Standard Deviation (for 2010):
=STDEV(B3:B30) = 65.7 ***[See Excel Figure, cell B33]***

Standard Deviation (for 2011):
=STDEV(C3:C30) = 62.2 ***[See Excel Figure, cell C33]***

4. Average spread is the average of the two standard deviations in step 3 above:

=AVERAGE(B33:C33) = 64.0 ***[See Excel Figure, cell D34]***

5. Overall Effect size is equal to the difference between the two mean scores (post-test and pre-test) divided by the average Standard Deviation. Therefore we need to divide the result in step 2) by the result in step 4) above:

=D32/D34* = 58.7** ÷ 64.0 = 0.92 ***[See Excel Figure, cell E35]

6. Individual student effect size is equal to the difference between the individual student post-test and pre-test score divided by the average Standard Deviation for the class:

***=D3/D34 (for student 17)* = -16** ÷ 64.0 = -0.25
[See Excel Figure, cell E3]

***… D30/D34 (for student 13) =* 126** ÷ 64.0 = 1.97
[See Excel Figure, cell E30]

*For assessments that measure change over 2 years, it is necessary to divide the effect size figure by 2 to approximate yearly growth, particularly when comparisons are made with other yearly based effect size figures (e.g. Appendix 2).

**All cell locations in Excel have a referencing system that are needed for calculating formulas. For example, Student Id:17 scored 551 in 2010 & in 2011 scored 535. The cell location e.g. cell 'B3' refers to *row*3 and *column* B. locate 'B3' (i.e. row 3 & column B), the value is '551'.

Standard Deviation (SD)

SD is a measure of the spread of all individual student scores relative to the mean score. When comparing the SD for schools with the same mean score, a larger SD indicates a larger spread of scores (i.e. more lower and higher scores).

Is the effect size a real and accurate result?

To determine whether the effect size is a real result, a confidence interval may be used to describe the level of uncertainty (or error) of inferring the true value, but this calculation is not within scope of this paper. There are also measurement errors that can occur when assessments are not properly designed or due to differences in test administration. Effect size calculations are recommended for assessments that have high levels of validity and reliability (e.g. validated and research based standardised/norm-referenced assessments). These factors are a reminder that effect size is not a precise or absolute measure of 'true' impact resulting from an intervention, but an estimate only.

Draft

Appendix 2: Table of Effect sizes of Influences

The following table provides information about the large range of strategies and programs of learning and their influence on student achievement as measured by effect size. The research indicates that the majority of interventions and strategies have an influence or level of workability. It is recommended that this information be used by educators to further discuss, evaluate and question what might be able to be changed (i.e. low influences) or strengthened (i.e. high influences) as part of educational practice.

Table of selected effect sizes of influences on student achievement. *Source: Hattie, J., (2012). Visible Learning for Teachers, Maximising Impact on Learning. Pages 251-256. Routledge, Oxford: UK.*

HIGH INFLUENCES	Effect Size
How to develop high expectations for each student	1.44
Providing formative evaluation to teachers	0.90
Classroom discussion	0.82
How to provide better feedback	0.75
Teacher- student relationships	0.72
How to better teach meta- cognitive strategies	0.69
Vocabulary programs	0.67
How to accelerate learning	0.68
Teaching Study Skills	0.63
Teaching learning strategies	0.62
Ways to stop labelling students	0.61
Comprehension programs	0.60
MEDIUM INFLUENCES	
Direct instruction	0.59
Cooperative vs individualistic learning	0.59
Phonics instruction	0.54
Peer influences on achievement	0.53
Influence of home environment	0.52
Professional development on student achievement	0.51
Parental involvement	0.49
Early intervention	0.47
How to develop high expectations for each teacher	0.43
Integrated curricular programs	0.39
Computer – assisted instruction	0.37
Decreasing disruptive behaviour	0.34
Homework	0.29
Teaching test- taking and coaching	0.27
LOW INFLUENCES	
School finances	0.23
Individualized instruction	0.22
Reducing class size	0.21
Extra-curricular programs	0.19
Home-school programs	0.16
Ability group/ tracking/streaming	0.12
Male and female achievement differences	0.12
Student control over learning	0.04
Open vs traditional learning spaces	0.01
Retention (holding back a year)	-0.13

Draft