Cambridge Assessment International Education

Evaluation – why it is important, but so difficult, for us to understand what works in education

Lee Davis
Deputy Director, Education

Sept 2018

Contents

1. Warm-up activity
2. The concept of meta-analysis
3. Critique of its use in education
4. Understanding impact for yourselves

What has the greatest impact on student achievement?

1. Revise las influencias
2. Decida si cada influencia tiene un impacto alto, medio o bajo
3. Por favor, trabaja en parejas

Los bebés deben dormir…?

Los bebés deben dormir…?

“The scandalous failure of scientists to accumulate scientifically!”

(Chalmers 2005)

Problem

Summaries of research produced by experts and peer-reviewed, but…

1. Lack of transparency
2. Bias

Response

- The *Cochrane Collaboration* (1990s) for science and medicine
- The *Campbell Collaboration* (2000) in social sciences
- Education Endowment Foundation (2011)

Response

Meta-analysis – the statistical analysis of a large collection of research findings, with the purpose of integrating these findings such that we can draw conclusions about the effectiveness of a particular intervention on students.

Meta-analysis method

- Define the research area, eg collaborative learning,
- Formulate a search strategy, eg online databases only, journal articles
- Define the inclusion criteria, eg randomised control trials, comparison of teachers and their levels of experience, age range etc.

The purpose is to identify the evidence base and therefore provide a summary of what is known about the area within the limitations specified above.

Meta-analysis method

Critically, the studies must be quantitative in design so that we can measure the impact of a given intervention on student outcomes, eg test scores.

This produces an *effect size* for each intervention

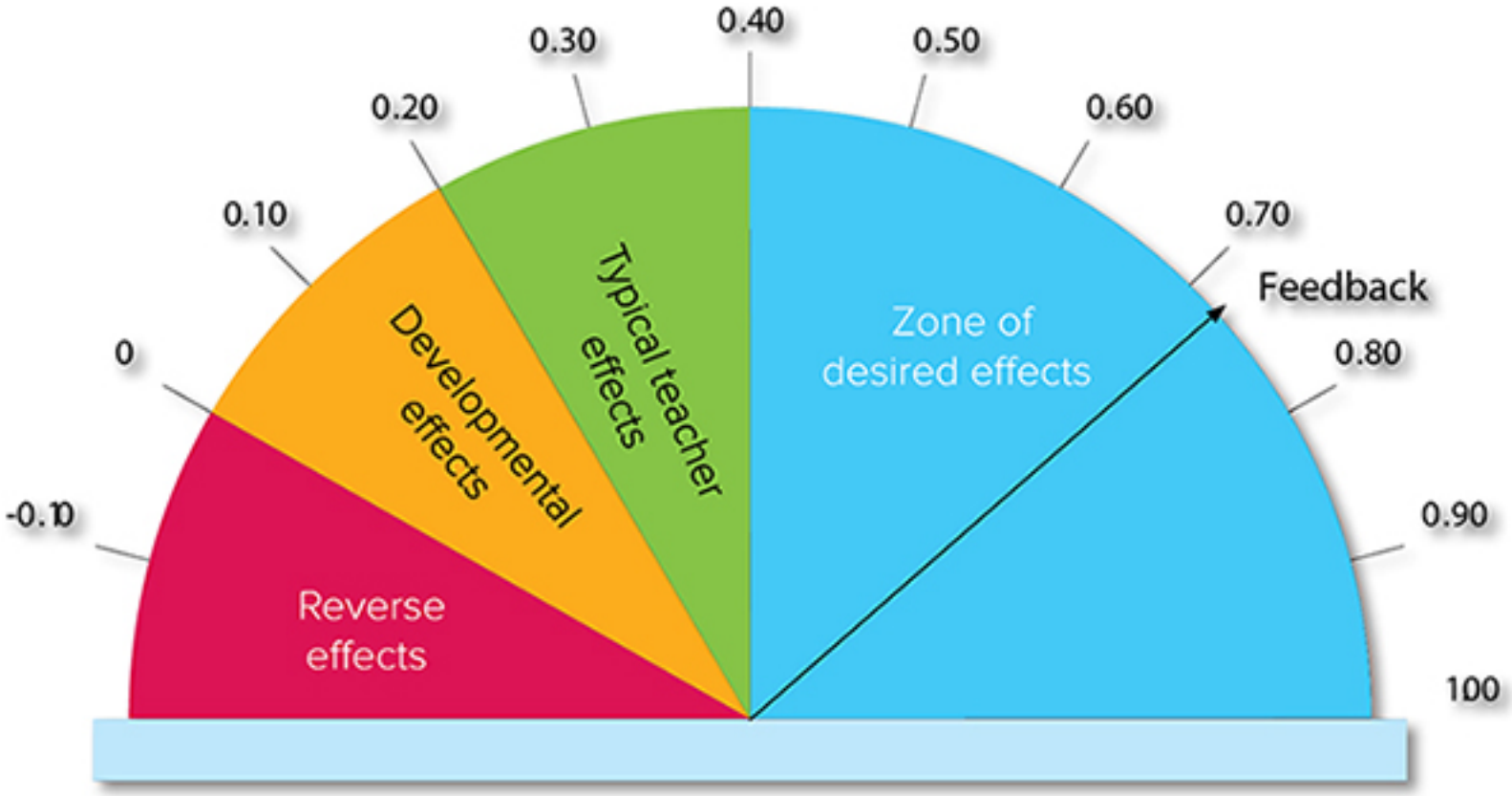
Calculating Effect Sizes

Definition

"The standardised mean difference between two groups."

$$\text{Effect Size} = \frac{[\text{Mean of Experimental Group}] - [\text{Mean of Control Group}]}{\text{Standard Deviation}}$$

Interpreting effect sizes



Respuestas

Influencia	Tamaño del efecto	Rank	Clasificación
Agrupar estudiantes de acuerdo a su habilidad	0.12	131	Low
Aceleración (por ejemplo, omitir un año)	0.68	15	High
Programas de comprensión de lectura	0.60	26	High
Concept mapping (Ayudando a los estudiantes a identificar las 'grandes ideas' dentro de un tema)	0.60	27	High
Aprendizaje cooperativo versus individualista	0.59	28	Medium
Instrucción directa	0.59	29	Medium
Feedback / retroalimentación	0.75	10	High
Género (logro masculino comparado con logro femenino)	0.12	133	Low
Ambiente en el hogar	0.52	44	Medium
Instrucción individual	0.22	109	Low
Influencia de los compañeros	0.53	41	Medium
Adaptación de la enseñanza con los estilos de aprendizaje de los alumnos	0.17	125	Low

Caveat Lector!

Problems

- Randomised control trials
- Intensity and duration of the intervention
- The *file drawer* problem
- Age of students
- Student outcome measures
- University laboratory problem

Randomised control trials

- Randomised control trials
  - Who gets the intervention and who does not?
  - Do teachers do what they are asked?
  - Does *ceteris paribus* really apply?

Intensity and duration of the intervention

Intensity and duration of the intervention

- Length of the intervention?
- How many teachers involved?
- How many students involved?
- Averaging effect sizes!
- The what! Eg reducing class sizes…

The *file drawer* problem

The *file drawer* problem

- Null hypothesis testing
- Larger effect sizes produce more statistically significant results
- Larger numbers of participants
- Lower thresholds on the *P*-value, eg <0.10
- Therefore studies that are published tend to overstate the effect of an intervention.
- In education, research has found that only 40% of studies are likely to produce a statistically significant result!
- So a lot of studies are confined to the *file drawer*.

The age of students

Question: would you expect to see greater variability, in terms of test scores, in the higher age ranges of students or in the lower age ranges?

Answer: variability is greater in the higher age ranges

- Higher the stdev the lower the Effect Size.
- Implications?
  - For the intervention?
  - For publishing?

Student outcome measures

Question: how do we measure student outcomes?

Answer:

1. Journal entries, book scrutiny, classroom tests (**immediate**)
2. Formal embedded assessment (**close**)
3. A different assessment of the same concept requiring some transfer (**proximal**)
4. Large-scale assessment from state or national curriculum framework (**distal**)
5. Standardised achievement tests (**remote**)

The laboratory problem

The generalisability of the study

- Eg Feedback

Conclusion

Using meta-analyses to determine the effectiveness of any given intervention in education is “almost useless”! (Wiliam 2016)

So what can we do?

- Read and interpret the research literature with care
- Use sites such as EEF for support
- Teachers can become their own evaluators using the effect size methodology

What was the impact on students?

Student name	Score pre intervention	Score post intervention
Michael	11	15
Sanjay	9	16
Peter	8	17
Ruchira	12	13
Anne	8	8
Nivedita	11	14

What was the impact on students?

Student name	Score pre intervention	Score post intervention
Michael	11	15
Sanjay	9	16
Peter	8	17
Ruchira	12	13
Anne	8	8
Nivedita	11	14
Average	**9.83**	**13.83**

What was the impact on students?

Student name	Score pre intervention	Score post intervention
Michael	11	15
Sanjay	9	16
Peter	8	17
Ruchira	12	13
Anne	8	8
Nivedita	11	14
Average	**9.83**	**13.83**
STDEV	**3.21**	

What was the impact on teachers?

Student name	Score pre intervention	Score post intervention
Michael	11	15
Sanjay	9	16
Peter	8	17
Ruchira	12	13
Anne	8	8
Nivedita	11	14
Average	**9.83**	**13.83**
STDEV	**3.21**	
Effect size for the group		**1.2**

What was the impact on students?

Student name	Score pre intervention	Score post intervention	Individual effect size
Michael	11	15	1.24
Sanjay	9	16	2.18
Peter	8	17	2.81
Ruchira	12	13	0.31
Anne	8	8	0
Nivedita	11	14	0.93
Average	**9.83**	**13.83**	
STDEV	**3.21**		
Effect size for the group		**1.2**	

Conclusions

- We need to share good practice as much as possible
- But take care that you interrogate the evidence critically
- Become evaluators of your own practice
  - It increases visibility
  - It encourages teachers to focus on what works
  - It encourages teachers to collaborate
  - It increases assessment literacy
  - It increases accountability

Thank you!