

# Is there any prospect of an agreed method for calculating effect sizes?

Greg Brooks

Emeritus Professor of Education,  
University of Sheffield

Presentation to

‘Challenges and potential solutions in the selection of appropriate outcome measures for RCTs’ workshop,

University of York, 8 September 2017

# References

... are on the last slide

# Ideally, there would be ...

- 1) an agreed method and formula for calculating effect sizes (ES),
- 2) and an agreed justification for them,

all these things would be known,

and everyone would just reference them

## Currently, however, ...

agreement extends only as far as the notion that an ES is the difference in mean scores between the treatment and control groups divided by an s.d.

– but which mean scores and which s.d?

# Variations

Numerator:

post-test means / mean gain scores

Denominator:

pooled pre-test s.d's / pooled post-test s.d's /  
pooled s.d's of the gain scores

# Tacit agreement?

'Don't' ask, don't tell'

BUT ...

# Deeply unsatisfactory

... because

- 1) not sufficient to cite statistics program used or name (e.g.) Cohen's  $d$  or Hedges'  $g$ , since neither specifies which mean scores or s.d.'s to use
- 2) if effect sizes are calculated using different methods, results will be strictly incommensurate, and combining such mixed effect sizes in meta-analyses may vitiate the conclusions drawn
- 3) this in turn may mean that educationally/socially useful findings are missed or (worse) unsound findings are promulgated

# Example 1

Ehri *et al.* (2001: 401):

‘The formula ... consisted of the mean of the treatment group minus the mean of the control group divided by a pooled standard deviation.’

National Reading Panel (2000: 1-10):

$$(M_t - M_c) / 0.5(sd_t + sd_c) \approx \text{Cohen's } d$$

But - which means/s.d's?

- denominator as arithmetic  $\bar{x}$  of s.d's not best practice



## Example 2

Gorard *et al.* (2015):

Group mean scores widely different at pre-test and control group still ahead at post-test, therefore effect size calculated using difference in gain scores, but over **pooled s.d's of gain scores:  $d = 0.25$**

Effect size calculated using difference in gain scores over **pooled post-test s.d's:  $d = 0.19$**

## Example 3

Of 12 systematic reviews of phonics schemes,

only one corrected for small Ns

(and then inconsistently,

and the review is rubbish for other reasons anyway)

# Second-best precautionary approach?

Insist that

- every trial report and systematic review states explicitly, and justifies, how the reported ESs were calculated, including showing the relevant formula
- this applies also to journal articles

Thank you

**Controversies?**

[g.brooks@sheffield.ac.uk](mailto:g.brooks@sheffield.ac.uk)

# References

Ehri, L.C., Nunes, S.R., Stahl, S.A. & Willows, D.M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, 71, 3, 393-447.

Gorard, S., Siddiqui, N. and See, B.H. (2015). *Fresh Start. Evaluation Report and Executive Summary*. London: Education Endowment Foundation.

National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington DC: National Institute for Child Health and Human Development Clearinghouse.