Synthesis of research findings

- How do we use findings from previous research?
- What counts as evidence?
- How do we ensure it is cumulative?
- How do we know it is applicable?

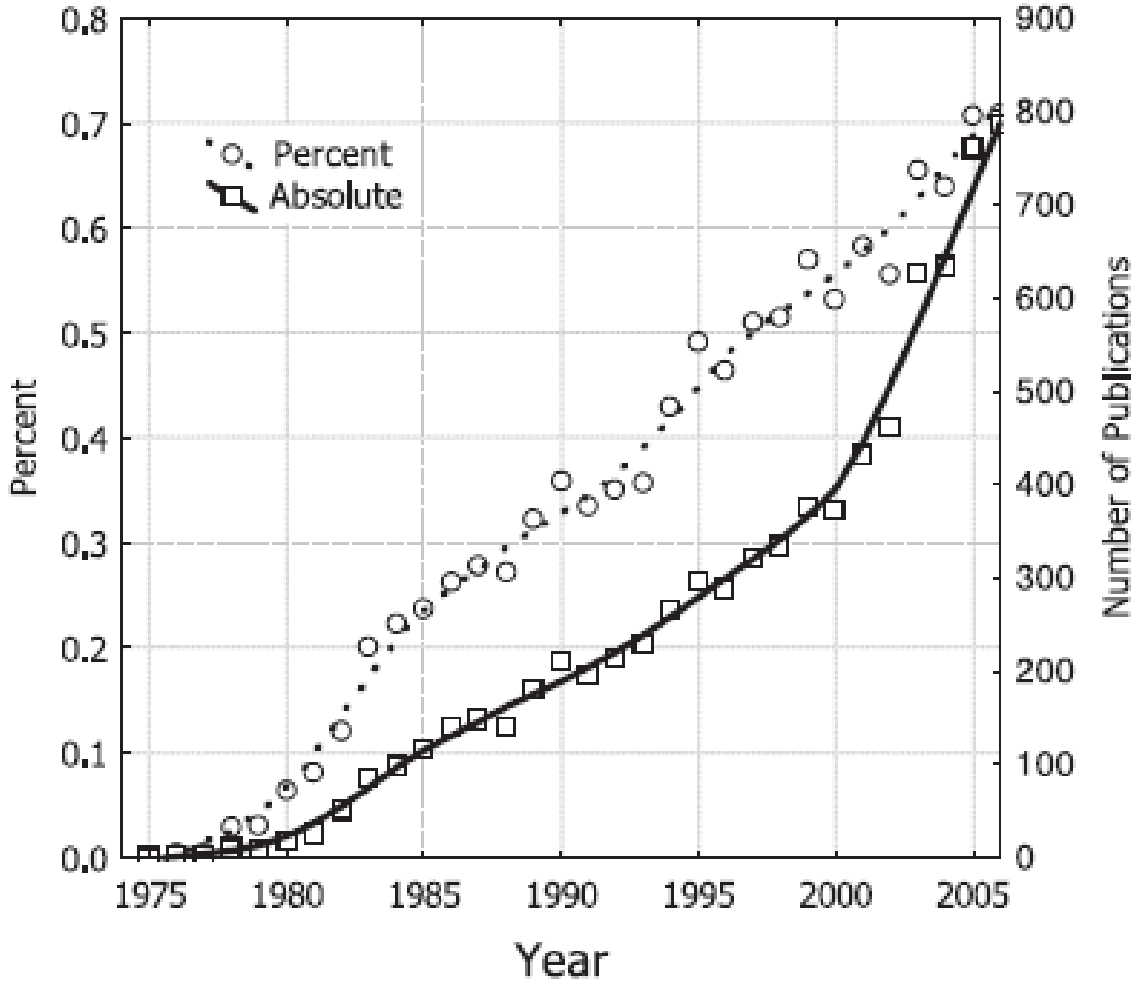


Figure 1. The absolute number and percentage of publications on meta-analysis in the database PsycINFO in the last 30 years.

Scenario

Imagine you're in a school governors' meeting and that you are discussing the school's homework strategy.

Someone waves around a review of research which they found on the internet which says has found that children should not be set more than half an hour of homework per night.

What questions would you have about how the review was done in order to know how it can help decisions about a homework strategy for the school?

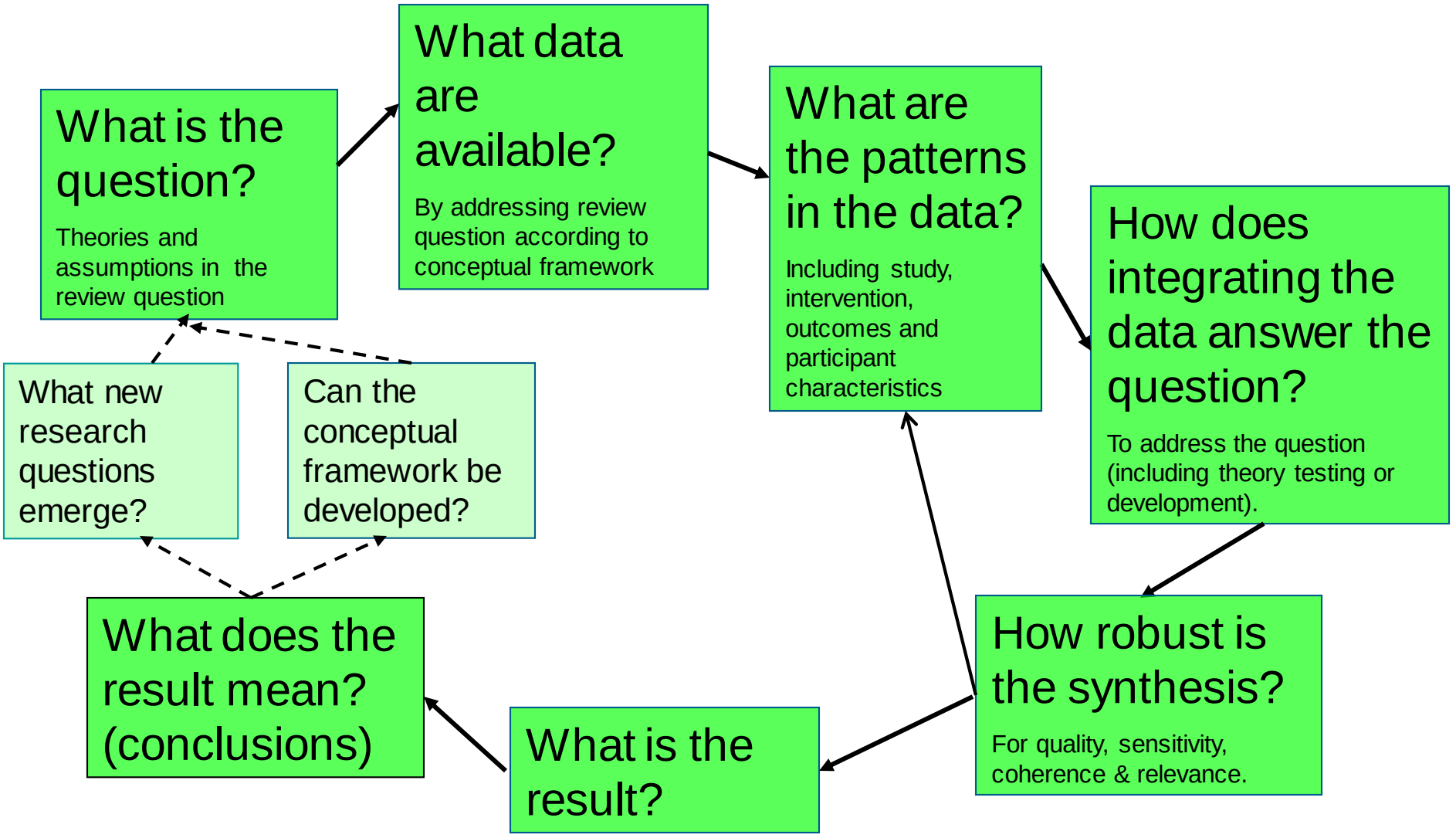
Key issues about reviews and evidence

- Applicability of the evidence to the question
  - Breadth
  - Scope
  - Scale
- Robustness of the evidence
  - Research quality

Stages of synthesis

- Stages in the conduct of most reviews or syntheses:
  - Review question and conceptual framework
  - Initial organization of data
  - Identifying and exploring patterns in the data
  - Integration of the data (synthesis)
  - Checking the synthesis
- *But* the process should not be seen as linear

Stages of synthesis



Cooper, H.M. (1982) Scientific Guidelines for Conducting Integrative Research Reviews *Review Of Educational Research* 52; 291
See also: Popay et al. (2006) *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews.* Lancaster: Institute for Health Research, Lancaster University.
http://www.lancs.ac.uk/fass/projects/nssr/research.htm

What is a systematic review?

- Some labels include ...
  - research synthesis,
  - research review,
  - systematic review,
  - integrative review
  - quantitative review, and
  - meta-analysis.
- *The term "meta-analysis" sometimes refers only to quantitative summaries and sometimes more broadly.*

Systematic reviewing

- Key question
- Search protocol
- Inclusion/exclusion criteria
- Coding and mapping
- In-depth review (sub-question)
- Techniques for systematic synthesis

What is the question?

What data are available?

What patterns are in the data?

How robust is the synthesis?

What are the results?

Advantages

- uses explicit, replicable methods to identify relevant studies, then
- uses established or transparent techniques to analyze those studies; and
- aims is to limit bias in the identification, and evaluation of studies and in the integration or synthesis of information applicable to a specific research question.

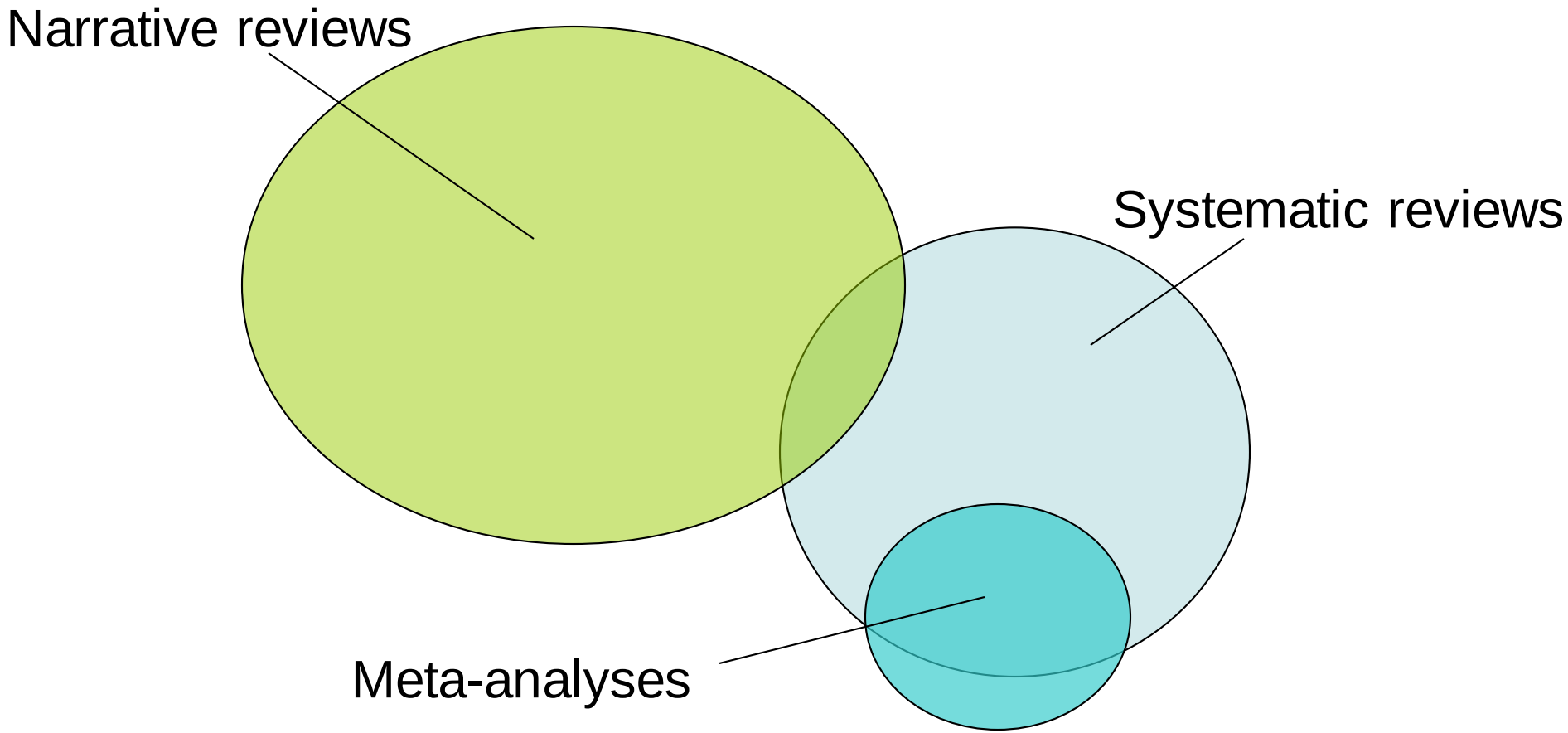
Underpinning bias in systematic reviews?

- Research and policy focus
- Specific reviews to answer particular questions
  - What *works*? - impact and effectiveness research with a resulting tendency to focus on quantitative and experimental designs

Meta-analysis as synthesis

- Quantitative data from
  - Experimental research studies
  - Correlational research studies
- Methodological assumptions from quantitative approaches (both epistemological and mathematical)

Literature reviews - conceptual relations



Meta-analysis or quantitative synthesis

- Synthesis of quantitative data
  - Cumulative
  - Comparative
  - Correlational
- "Surveys" educational research (Lipsey and Wilson, 2001)

Origins

1952: Hans J. Eysenck concluded that there were no favorable effects of psychotherapy, starting a raging debate which 25 years of evaluation research and hundreds of studies failed to resolve

1978: To prove Eysenck wrong, Gene V. Glass statistically aggregated the findings of 375 psychotherapy outcome studies

- Glass (and colleague Smith) concluded that psychotherapy did indeed work "the typical therapy trial raised the treatment group to a level about two-thirds of a standard deviation on average above untreated controls; the average person received therapy finished the experiment in a position that exceeded the 75th percentile in the control group on whatever outcome measure happened to be taken" (Glass, 2000).
- Glass called the method "meta-analysis"

(adapted from Lipsey & Wilson, 2001)

Historical background

- Underpinning ideas can be identified earlier:
  - K. Pearson (1904)

  Averaged correlations for typhoid mortality after inoculation across 5 samples
  - R. A. Fisher (1944)

    "When a number of quite independent tests of significance have been made … although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance" (p. 99).

    Source of the idea of cumulating probability values
  - W. G. Cochran (1953)

    Discusses a method of averaging means across independent studies

    Set out much of the statistical foundation for meta-analysis (e.g., inverse variance weighting and homogeneity testing)

(adapted from Lipsey & Wilson, 2001 and Hedges, 1984)

Meta-analysis

- Key question
- Search protocol
- Inclusion/exclusion criteria
- Coding
- Statistical exploration of findings
  - Mean and range
  - Distribution
  - Sources of variance
  - ‘Sensitivity’

What is the question?

What data are available?

How robust is the synthesis?

Intervention research

- Usually evaluation of policies, practices or programmes
- Usually based on experiments (RCTs, quasi-experimental designs)
- Answering impact questions
  - Does it work?
  - Is it better than…?

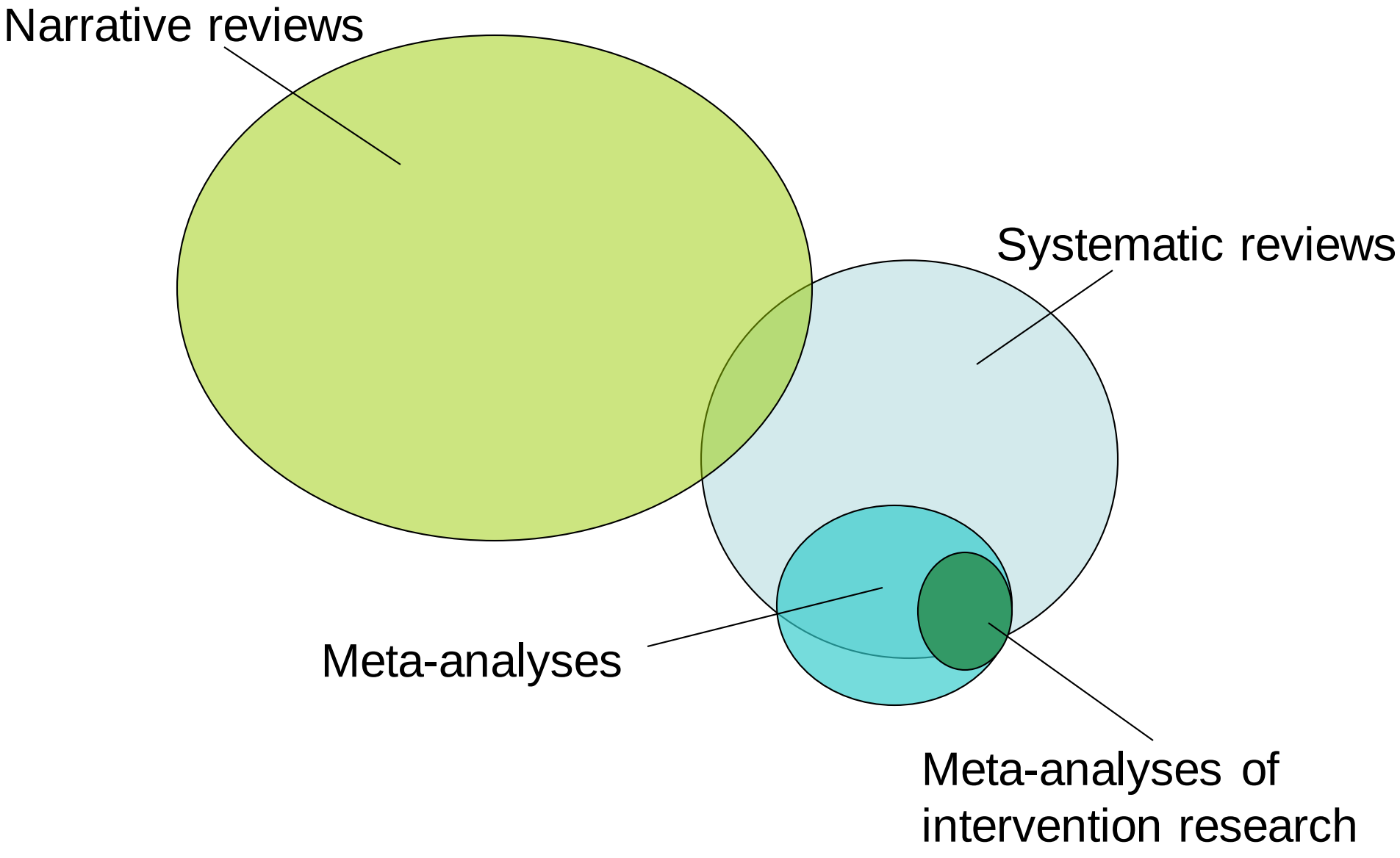
Impact questions

- Causal
  - Does X work better than Y?
    - Homework intervention studies
- Not correlational
  - Rather than associational
    - Do schools with homework do better?

Kinds of questions…

- Identify an area of research you are interested in
- Discuss what kind of questions could be answered by
  a) Interventions
  b) Correlational studies

Literature reviews - conceptual relations



Comparing quantitative studies

- The need for a common measure across research studies
  - Identifying a comparable measure
  - Using this effectively
  - Interpreting this appropriately

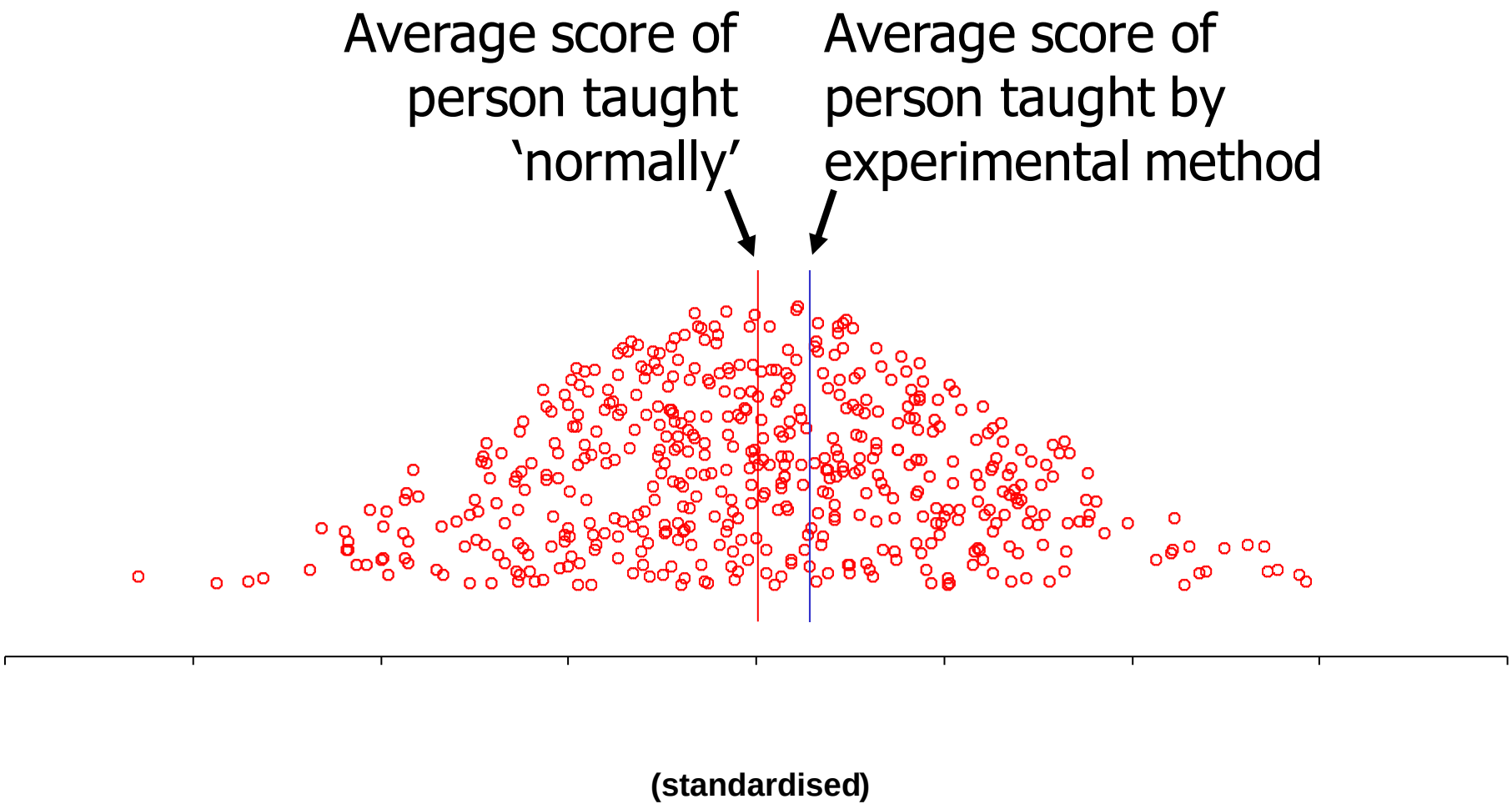
Significance versus effect size

- Traditional test is of statistical ‘significance’
- The difference is unlikely to have occurred by chance
  - However it may not be:
    - Large
    - Important, or even
    - Educationally ‘significant’

The rationale for using effect sizes

- Traditional reviews focus on statistical significance testing
  - Highly dependent on sample size
  - Null finding does not carry the same "weight" as a significant finding
- Meta-analysis focuses on the ***direction*** and ***magnitude*** of the effects across studies
  - From "*Is there a difference?"* to "*How big is the difference?"*
  - Direction and magnitude represented by "***effect size***"

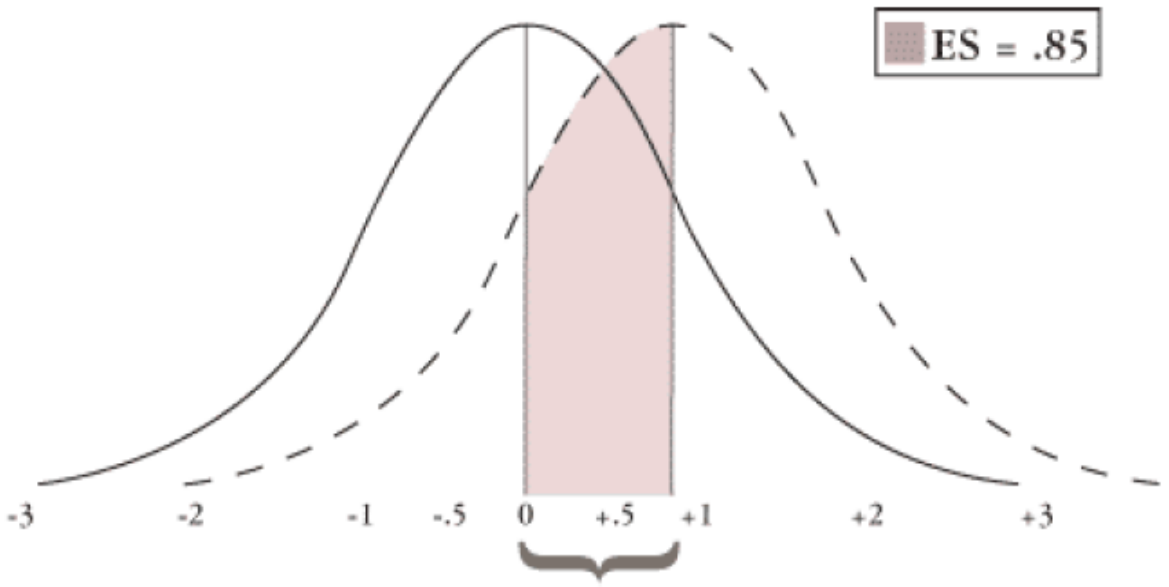
Effect size


Average score of person taught 'normally'
Average score of person taught by experimental method
(standardised)

$$\text{Effect size} = \frac{\text{Mean of experimental group} - \text{Mean of control group}}{\text{Standard deviation}}$$

Effect size is the difference between the two groups, relative to the standard deviation

Effect size



Percentile Point Gain = 30

From: Marzano, R. J. (1998) A Theory-Based Meta-Analysis of Research on Instruction. Aurora, Colorado, Mid-continent Regional Educational Laboratory. Available at: http://www.mcrel.org:80/topics/products/83/ (accessed 2/9/08).

Effect size

- Comparison of impact
- Same AND different measures
- Significance vs effect size
  - *Does it work?* vs *How* well *does it work?*

Wilkinson, L., & APA Task Force on Statistical Inference. (1999) Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

Effect sizes

- Standardised way of looking at difference
  - Different methods for calculation
    - Correlational (e.g. Pearson's r)
    - Odds ratio (binary/dichotomous outcomes)
    - Standardised mean difference
      - Difference between control and intervention group as proportion of the dispersion of scores

Effect size

- The difference between the two means, expressed as a proportion of the standard deviation
- ES =(M_e – M_c) / SD
- Issues
  - Which standard deviation?
  - Statistical significance?
  - Margin of error?
  - Normal distribution?
  - Restricted range
  - Reliability

Main approaches

Cohen's d

(but which SD?)

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s},$$

Glass's Δ

(sd of control)

$$\Delta = \frac{\bar{x}_1 - \bar{x}_2}{s_2}$$

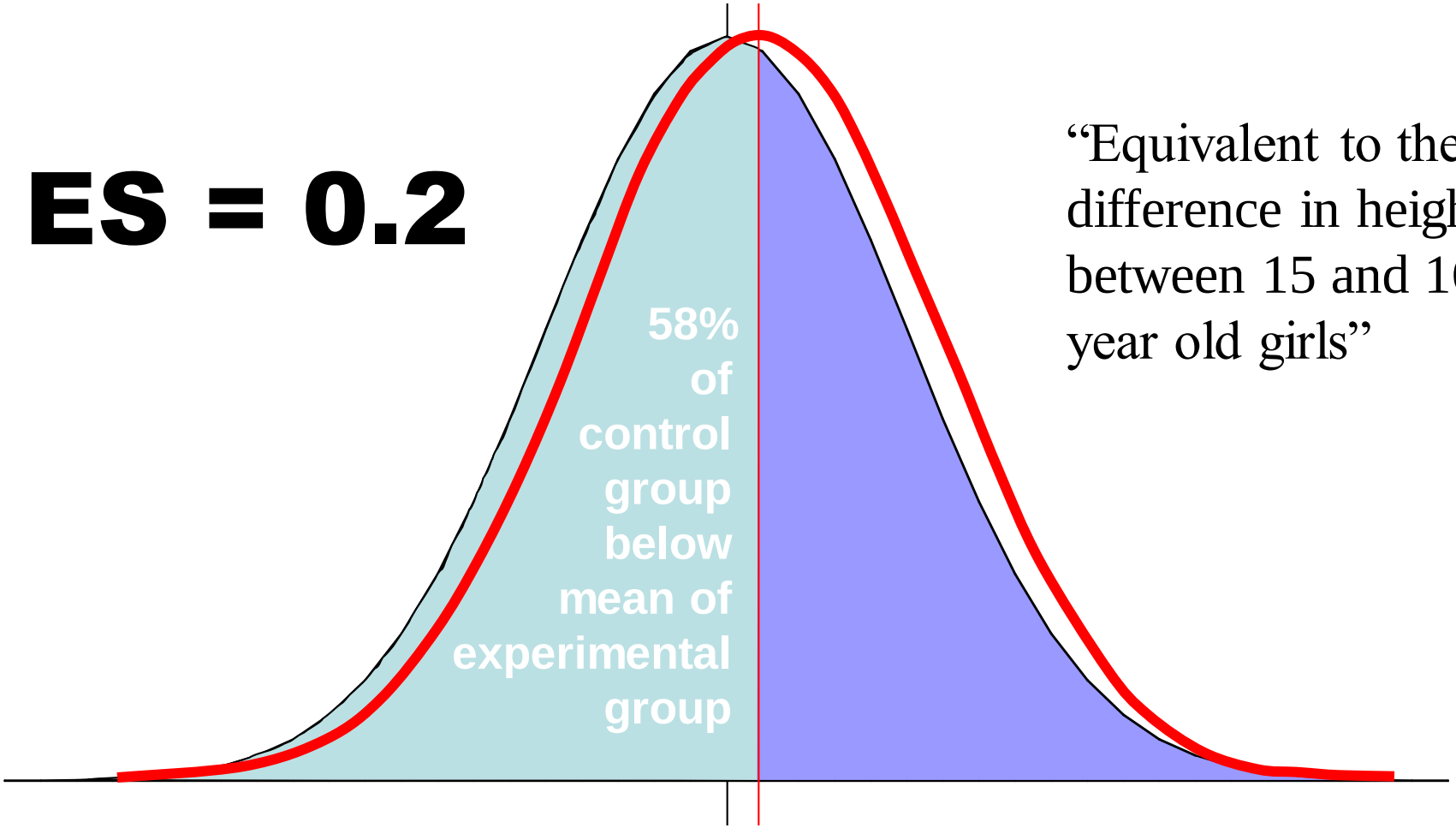
Hedges' g

(weighted for sample size)

$$g = \frac{\bar{x}_1 - \bar{x}_2}{s^*} \qquad s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Examples of Effect Sizes:

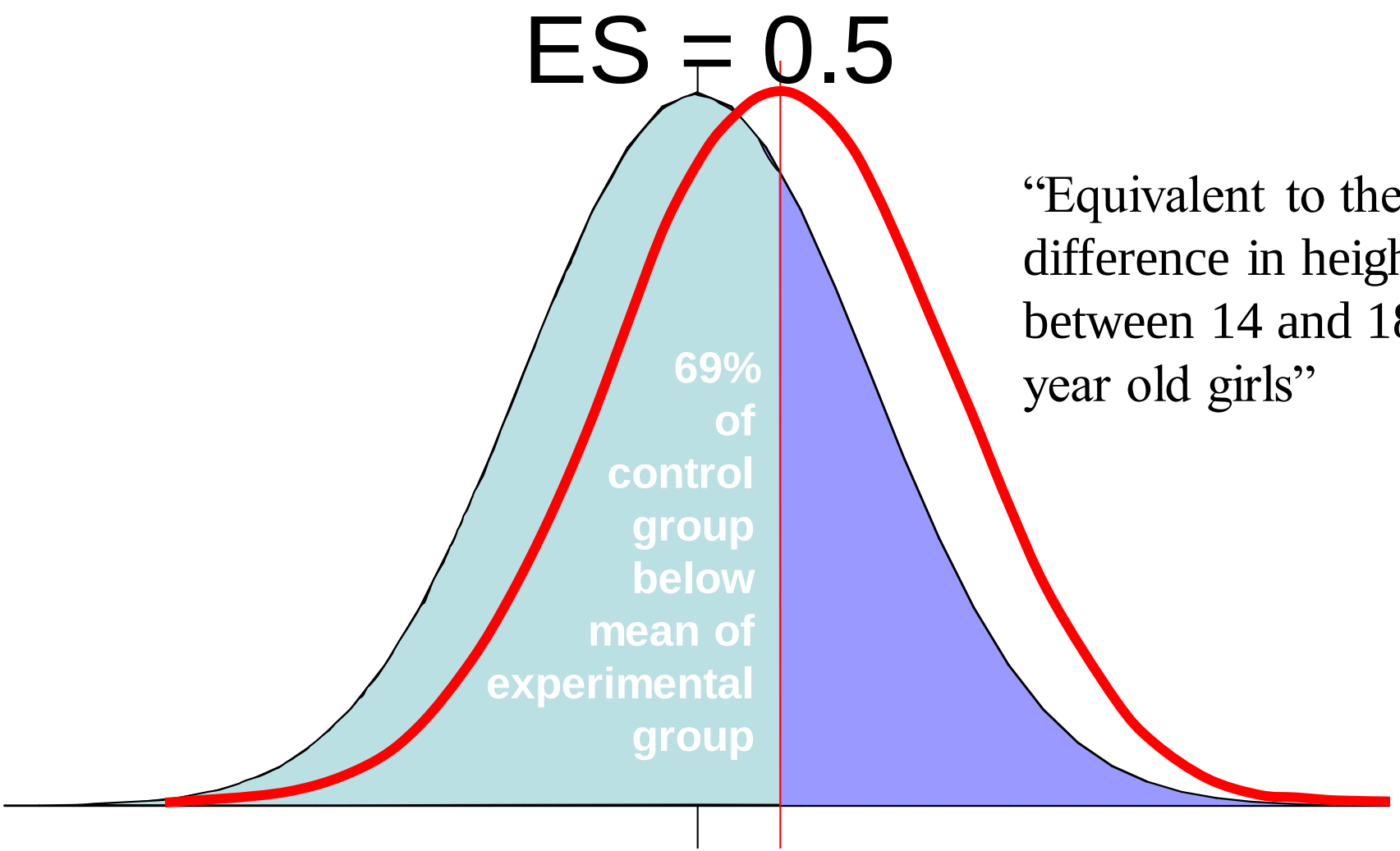
ES = 0.2



"Equivalent to the difference in heights between 15 and 16 year old girls"

Probability you could guess which group a person was in = 0.54

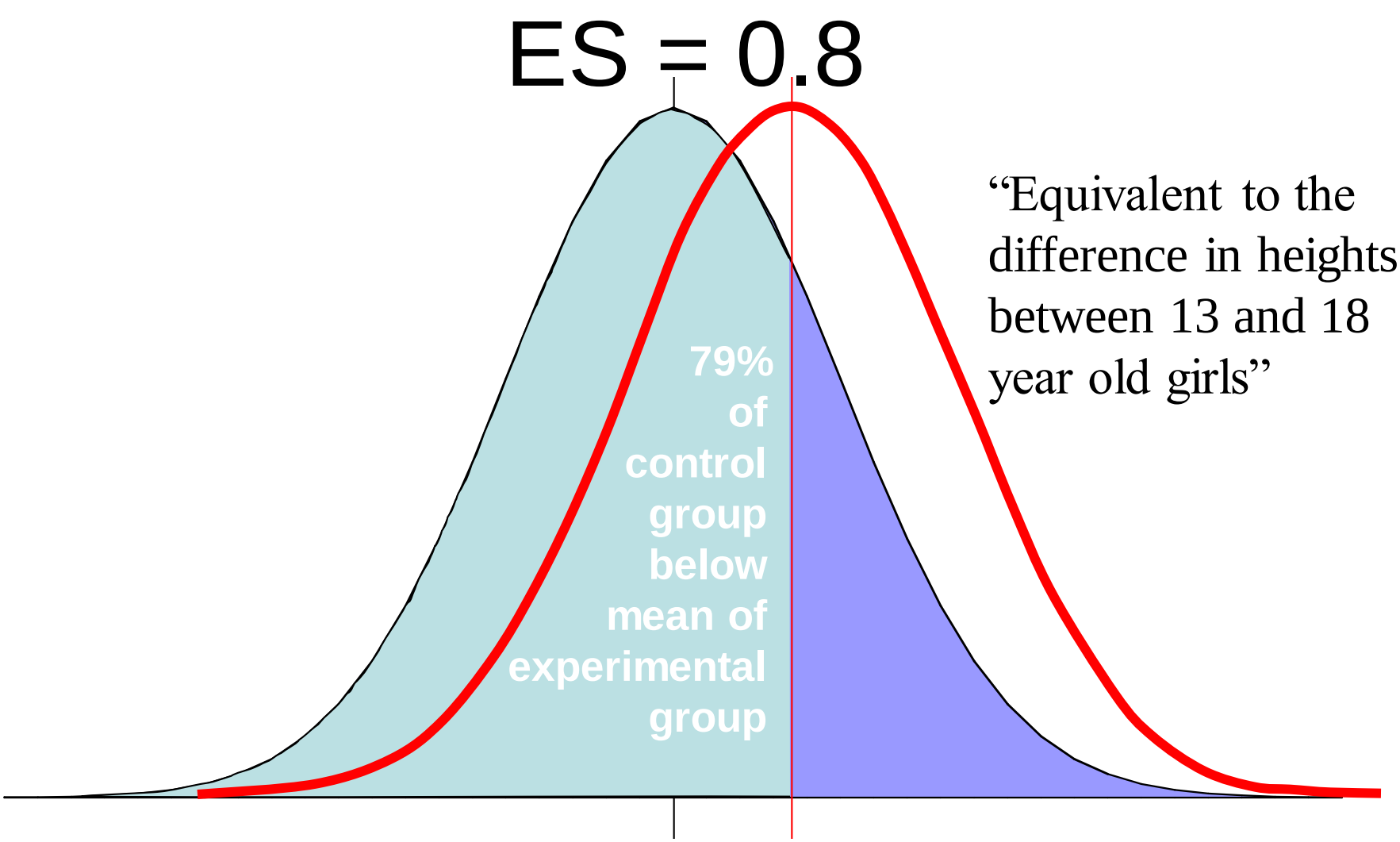
Change in the proportion above a given threshold:
from 50% to 58% *or* from 75% to 81%



"Equivalent to the difference in heights between 14 and 18 year old girls"

Probability you could guess which group a person was in = 0.60

Change in the proportion above a given threshold:
from 50% to 69% *or* from 75% to 88%



Probability you could guess which group a person was in = 0.66

Change in the proportion above a given threshold:
from 50% to 79% *or* from 75% to 93%

Rank (or guess) some effect sizes…

Learning styles

ICT/Educational technology

Homework

Providing feedback

Direct instruction

“small” ≤ 0.2 “medium” 0.21 - 0.79 or “large” ≥ 0.8

Rank order of effect sizes

0.79 Providing feedback (Hattie & Timperley, 2007)

0.6 Direct instruction (Sipe & Curlette, 1997)

0.37 ICT/Ed Tech (Hattie, 2008)

0.29 Homework (Hattie, 2008)

0.15 Learning styles (Kavale & Forness, 1987; cf Slemmer 2002)

Interpreting effect sizes

- a “small” effect may be important in an intervention which is cheap or easy to implement
- a “small” effect may be meaningful if used across an entire population (prevention programs for school children)
- “small” effects may be more achievable for serious or intractable problems
- *but Cohen’s categories correspond with the broad distribution of effects across meta-analyses found by Lipsey and Wilson (1993), Sipe and Curlette (1997) and Hattie (2008)*

Confidence intervals

- Robustness of the effect
  - Shows the range within which a presumed actual effect is likely to be
    - Smaller studies - larger confidence intervals
    - Larger studies - smaller confidence intervals
  - If a confidence interval includes zero, the intervention is not significantly different statistically from the control
  - Does not avoid issues of bias in the synthesis

Effectiveness of Volunteer Tutoring Programs

Study	Outcome	Hedges Õ g	CI lower	CI upper	Sample (A, B)
Allor 2004	Combined	0.57*	0.10	1.04	61 25
Baker 2000	Combined	0.40	-0.02	0.83	43 41
Cobb 2000	Combined	0.66	-0.25	1.57	
Cook 2001.1	RG-WRAT3	0.24	-0.51	0.99	12 14
Cook 2001.2	RG-WRAT3	0.23*	0.11	0.35	11 6
Erion 1994	RA-Reading fluency	0.43	-0.35	1.22	12 12
Mayfield 2000	Combined	0.23	-0.27	0.73	31 29
McKinney 1995	RG-Stanford Reading	0.06	-0.52	0.64	20 24
Mehran 1988	Combined	0.47	-0.05	1.00	28 28
Miller 1994	RG-GORT-D	0.06	-0.51	0.63	23 23
Morris 1990.1	Combined	0.51	-0.16	1.18	17 17
Morris1 1990.2	Combined	0.58	-0.19	1.34	13 13
Nielson 1992	RC-Stanford Reading	0.28	-0.31	0.88	29 17
Powell-Smith 2000	Combined	-0.22	-0.90	0.45	24 12
Pullen 2004	Combined	0.54	-0.04	1.11	23 24
Rimm-Kaufman 1999	Combined	0.05	-0.55	0.64	21 21
Vadasy 2000	Combined	0.83*	0.24	1.43	23 23
Vadasy 1997a	Combined	0.51	-0.15	1.17	17 18
Vadasy 1997b	Combined	0.28	-0.33	0.89	20 20
Weiss 1989	Combined	-0.20	-1.11	0.71	9 8
	Overall	**0.30***	**0.18**	**0.42**	

Adapted from Ritter *et al*. (2006) p 38.

Confidence intervals

- By convention set at 95% level
  - 95 times out of 100 the population effect will be within the range of the confidence interval (in the context of estimation and assuming the same population)
  - Allows us to look at statistically non-significant results
  - Is a large effect with a wide confidence interval the same as a small effect and a narrow confidence interval?

Some recent findings from meta-analysis in education

Bernard *et al.* 2004

- Distance education and classroom instruction - 232 studies, 688 effects - wide range of effects ('heterogeneity'); asynchronous DE more effective than synchronous.

Pearson *et al.* 2005

- 20 research articles, 89 effects 'related to digital tools and learning environments to enhance literacy acquisition'. Weighted effect size of 0.49 indicating technology can have a positive impact on reading comprehension.

Klauer & Phye 2008

- 74 studies, 3,600 children. Training in inductive reasoning improves academic performance (0.69) more than intelligence test performance (0.52).

Gersten *et al.* 2009

- Maths interventions for low attainers. 42 studies ES ranging from 0.21-1.56. Teaching heuristics and explicit instruction particularly beneficial.