Common Effect Sizes in Experimental Designs

The “Big Ones”:

- **Continuous outcome family**: Mean Difference, Cohen’s d, Hedges’ g
- **Dichotomous outcomes family**: Relative Risk, Odds Ratio, Risk Difference
- **Time-to-event family**: Hazard Ratio, Incidence Rate Ratio

Effect Sizes for Continuous Outcomes

Between-Group Standardized Mean Difference

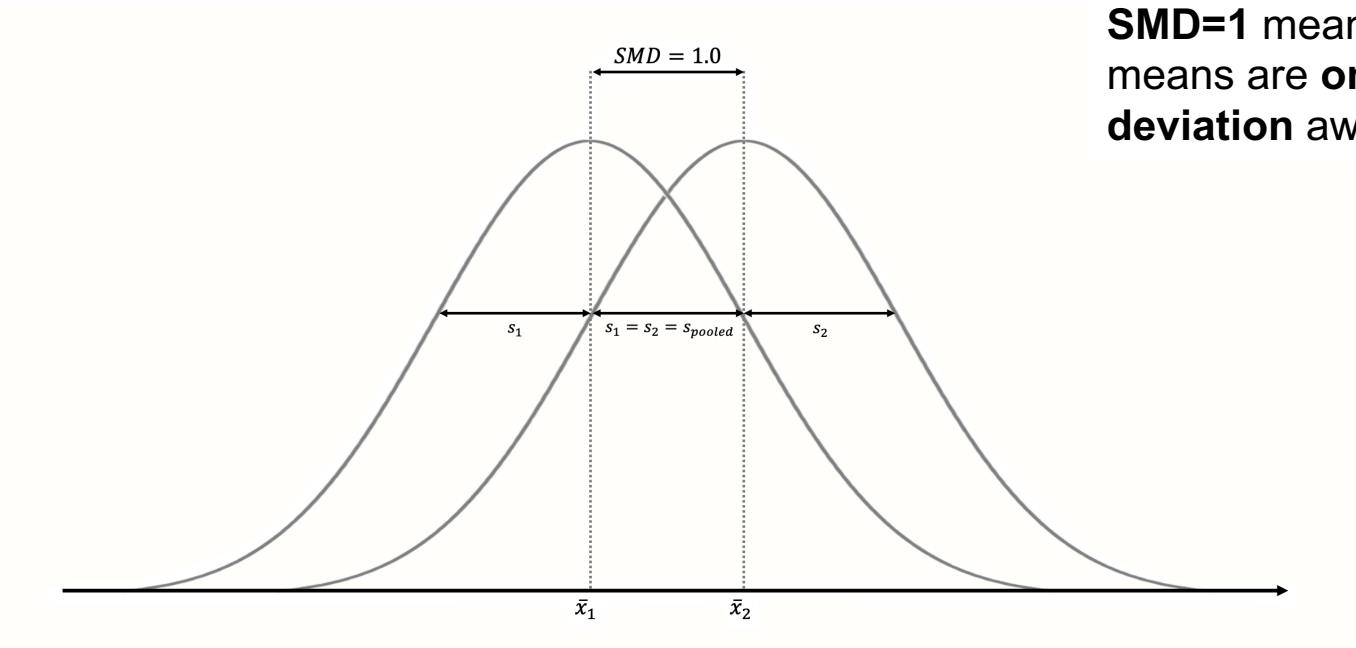
- Difference in means between two independent groups, standardized by the pooled standard deviation s_{pooled}.
- Often called **Cohen's** d after psychologist and statistician Jacob Cohen.

$$\text{SMD} = \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}} \qquad s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

$$\text{SE}_{\text{SMD}} = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{\text{SMD}^2}{2(n_1 + n_2)}}$$

Effect Sizes for Continuous Outcomes

Between-Group Standardized Mean Difference



SMD=1 means that the two group means are **one sample standard deviation** away from each other

Effect Sizes for Continuous Outcomes

Between-Group Standardized Mean Difference

Standardized mean differences are often interpreted using the conventions by Cohen (1988):

- SMD ≈ 0.20: small effect.
- SMD ≈ 0.50: moderate effect.
- SMD ≈ 0.80: large effect.

These are rules of thumb at best. It is usually much better to interpret standardized mean differences based on their “real-life” implications.

For many serious diseases, even a very small statistical effect can still have a huge impact on the population level.

Effect Sizes for Continuous Outcomes

Between-Group Standardized Mean Difference

The SMD has been found to have an upward bias when the sample size of a study is small (Hedges 1981).

This small sample bias means that SMDs systematically overestimate the true effect size when the total sample size of a study is small.

This small-sample bias is corrected for when we calculate **Hedges' g**:

$$g = SMD \times \left(1 - \frac{3}{4n - 9}\right)$$

Effect Sizes for Continuous Outcomes

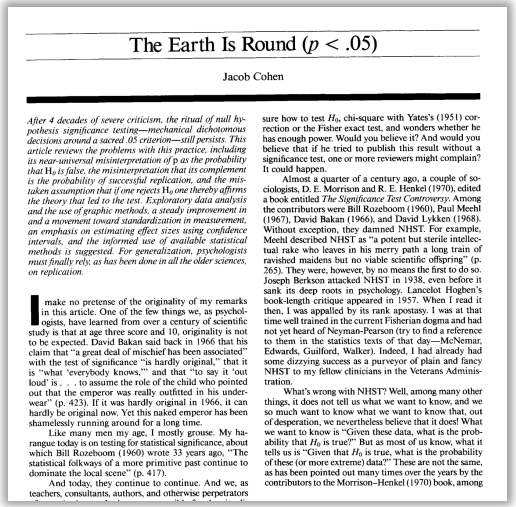
The Perils of Standardization

Standardization allows us, at least in theory, to compare the strength of an effect observed in different studies; even if these studies did not use the same instruments to measure it.

However, the size of a particular study's SMD depends heavily on the variability of its sample

Imagine two identical studies conducted in populations with drastically different variances → the SMD value of both studies would differ greatly

→ **"The effect of A on B for me can hardly depend on whether I'm in a group that varies greatly […] or another that does not vary at all"** (Cohen, 1994)



The Earth Is Round ($p < .05$)

Jacob Cohen

After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including its near-universal misinterpretation of p as the probability that H_0 is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects H_0 one thereby affirms the theory that led to the test. Exploratory data analysis and the use of graphic methods, a steady improvement in and a movement toward standardization in measurement, an emphasis on estimating effect sizes using confidence intervals, and the informed use of available statistical methods is suggested. For generalization, psychologists must finally rely, as has been done in all the older sciences, on replication.

I make no pretense of the originality of my remarks in this article. One of the few things we, as psychologists, have learned from over a century of scientific study is that at age three score and 10, originality is not to be expected. David Bakan said back in 1966 that his claim that "a great deal of mischief has been associated" with the test of significance "is hardly original," that it is "what 'everybody knows,'" and that "to say it 'out loud' is . . . to assume the role of the child who pointed out that the emperor was really outfitted in his underwear" (p. 423). If it was hardly original in 1966, it can hardly be original now. Yet this naked emperor has been shamelessly running around for a long time.

Like many men my age, I mostly grouse. My harangue today is on testing for statistical significance, about which Bill Rozeboom (1960) wrote 33 years ago, "The statistical folkways of a more primitive past continue to dominate the local scene" (p. 417).

And today, they continue to continue. And we, as teachers, consultants, authors, and otherwise perpetrators sure how to test H_0, chi-square with Yates's (1951) correction or the Fisher exact test, and wonders whether he has enough power. Would you believe it? And would you believe that if he tried to publish this result without a significance test, one or more reviewers might complain? It could happen.

Almost a quarter of a century ago, a couple of sociologists, D. E. Morrison and R. E. Henkel (1970), edited a book entitled *The Significance Test Controversy*. Among the contributors were Bill Rozeboom (1960), Paul Meehl (1967), David Bakan (1966), and David Lykken (1968). Without exception, they damned NHST. For example, Meehl described NHST as "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" (p. 265). They were, however, by no means the first to do so. Joseph Berkson attacked NHST in 1938, even before it sank its deep roots in psychology. Lancelot Hogben's book-length critique appeared in 1957. When I read it then, I was appalled by its rank apostasy. I was at that time well trained in the current Fisherian dogma and had not yet heard of Neyman-Pearson (try to find a reference to them in the statistics texts of that day—McNemar, Edwards, Guilford, Walker). Indeed, I had already had some dizzying success as a purveyor of plain and fancy NHST to my fellow clinicians in the Veterans Administration.

What's wrong with NHST? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is "Given these data, what is the probability that H_0 is true?" But as most of us know, what it tells us is "Given that H_0 is true, what is the probability of these (or more extreme) data?" These are not the same, as has been pointed out many times over the years by the contributors to the Morrison–Henkel (1970) book, among

Effect Sizes for Dichotomous Outcomes

Risks:

$$p_{\text{treat}} = \frac{a}{a+b} = \frac{a}{n_{\text{treat}}}$$

$$p_{\text{control}} = \frac{c}{c+d} = \frac{c}{n_{\text{control}}}$$

	Event	No Event	
Treatment	a	b	n_{treat}
Control	c	d	n_{control}
	n_E	$n_{\neg E}$	

Risk Difference or **Absolute Risk Reduction:**

$$\text{RD} = \text{ARR} = p_{\text{treat}} - p_{\text{control}}$$

Risk Ratio or **Relative Risk:**

$$\text{RR} = \frac{p_{\text{treat}}}{p_{\text{control}}}$$

Effect Sizes for Dichotomous Outcomes

Odds:

$$\text{Odds}_{\text{treat}} = \frac{a}{b}$$

$$\text{Odds}_{\text{control}} = \frac{c}{d}$$

Odds Ratio:

$$\text{OR} = \frac{a/b}{c/d} = \frac{\text{Odds}_{\text{treat}}}{\text{Odds}_{\text{control}}}$$

	Event	No Event	
Treatment	a	b	n_{treat}
Control	c	d	n_{control}
	n_E	$n_{\neg E}$	

Effect Sizes for Dichotomous Outcomes

A Numeric Example

$$\text{OR} = \frac{a/b}{c/d} = \frac{30/70}{50/50} = 0.43$$

$$\text{RR} = \frac{a/(a+b)}{c/(c+d)} = \frac{30/100}{50/100} = 0.60$$

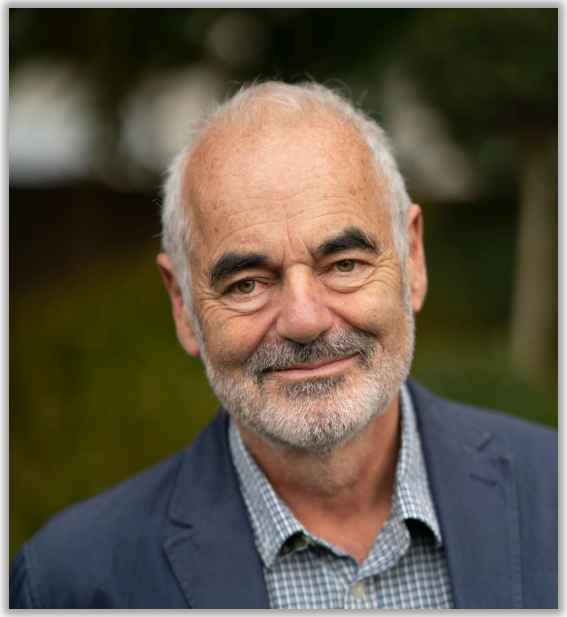
$$\text{RD} = \frac{a}{a+b} - \frac{c}{c+d} = 0.30 - 0.50 = -0.20$$

	Event	No Event	
Treatment	30	70	100
Control	50	50	100
	80	120	200

- The smaller the number of events, the more the OR and RR will agree.
- The RD is helpful to express effects on the **population level**, but depends heavily on the investigated population (and is thus less often used in meta-analysis)
- Many people (including clinicians) find the OR harder to understand than RR (or confuse ORs for RRs)

Effect Sizes for Dichotomous Outcomes

Video: “Bacon Sandwiches” by David Spiegelhalter



Effect Sizes for Dichotomous Outcomes

Non-Collapsibility of the Odds Ratio (Greenland, 2021)

Imagine the results of a depression psychotherapy RCT examining how many patients improve in the intervention (IG) and control group (CG):

Complete Sample

OR = 2.76

RR = 1.47

	CG	IG
Unchanged	137	73
Improved	136	200

No Antidepressants

OR = 3.95

RR = 1.29

	CG	IG
Unchanged	55	18
Improved	127	164

Antidepressants

OR = 5.96

RR = 4.00

	CG	IG
Unchanged	82	55
Improved	9	36

Effect Sizes for Dichotomous Outcomes

Non-Collapsibility of the Odds Ratio

- **This behavior is not caused by confounding, but by the numerical averaging failure of the odds ratio**

→ We must pay attention to ”which” odds ratio we extract from each study

→ Typically, we are only interested in the overall, or “marginal” odds ratio

→ In this case, it would be wrong to extract ORs from, e.g., logistic regression models in which certain covariates (e.g., antidepressant use, age, baseline symptoms etc.) were controlled for.

Effect Sizes for Dichotomous Outcomes

The Number Needed To Treat (NNT)

- Effect sizes such as Cohen's d, Hedges' g, odds ratios or relative risks are often difficult to interpret from a practical standpoint.
- How can we communicate what an effect means to patients, public officials, medical professionals, or other stakeholders?
- The NNT signifies how many additional patients must receive the treatment under study to prevent one additional negative event (e.g., relapse) or achieve one additional positive event
  - If NNT = 3, for example, we can say that three individuals must receive the treatment to avoid one additional relapse case; or that three patients must be treated to achieve one additional case of reliable symptom remission