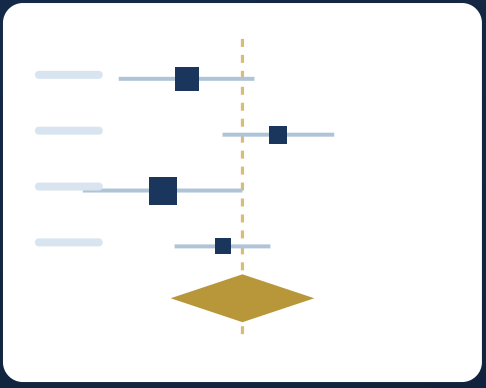
Research Gold

RESEARCH GOLD FIELD GUIDE

Choosing and Reading an Effect Size

The estimator that fits your design, the bias hiding in the default output, the interval that makes it defensible, and the reviewer objections that catch most submissions. Written for researchers who already report effect sizes and want to stop losing points on them.



Written by PhD methodologists researchgold.org

START HERE

The defensible question is not which effect size, but which one your design licenses

Most effect-size errors are not arithmetic. They are a mismatch between the estimator the software returned first and the design that generated the data, or a benchmark label read as if it were a finding. This guide is built around the choices a reviewer will actually probe.

PRINCIPLE

Pick the estimator from the data-generating design and the estimand you are willing to defend, not from the dialog box. A standardised mean difference and a response ratio can describe the same trial and answer different questions; an odds ratio and a risk ratio diverge precisely when the outcome is common. The estimator encodes an assumption. State it before you report the number.

The families, and the decision inside each

1. **Standardised differences between means**

   Cohen's d standardises the mean difference by a pooled standard deviation; Hedges' g multiplies d by the small-sample correction J = 1 - 3/(4(n1+n2) - 9), which removes a positive bias that is material below roughly 20 per group and negligible past 50. Glass's delta standardises by the control SD alone, and is the honest choice when the intervention plausibly changes the variance, not just the mean.

2. **Effects on binary outcomes**

   Risk ratio, odds ratio, and risk difference are not interchangeable. For cohort studies and trials, report the risk ratio or risk difference, because they track the quantity clinicians act on; reserve the odds ratio for case-control designs and logistic models, and remember it inflates away from the risk ratio once baseline risk exceeds about 10 per cent.

3. **Relationships between continuous variables**

   Pearson r assumes linearity and bivariate normality; its square is variance explained. For meta-analysis or any inference on r, work on the Fisher z scale, where the sampling variance is the stable 1/(n - 3). Use Spearman or a robust correlation when the relationship is monotone but not linear.

4. **Variance explained in models**

   In analysis of variance, eta squared is positively biased and partial eta squared can sum past one across effects and is not comparable across designs; omega squared or epsilon squared are the less biased reports. In regression, standardised coefficients let predictors share a scale, but they shift with the predictor mix, so report them alongside unstandardised coefficients.

THE SELECTOR

Match the estimator to the design, then to the estimand

Find the comparison you are making and the design that produced it. The right column is the one you can defend, with the caveat that most often trips a submission.

DESIGN TO ESTIMATOR, WITH THE CATCH

What you are comparing	Defensible estimator	The catch a reviewer looks for
Two group means	Hedges' g (small n), Cohen's d (large n)	Pooled vs control SD; report g, not d, below ~20/group
Means with unequal variances	Glass's delta or a robust d	Pooling a SD the intervention itself changed
A binary outcome, cohort or trial	Risk ratio or risk difference	Reporting an odds ratio as if it were a risk ratio
A binary outcome, case-control	Odds ratio	Risk ratio is not estimable from the design
Two continuous variables	Pearson r (Fisher z for inference)	Linearity and outliers; r on raw scale for pooling
Three or more group means	Omega squared or epsilon squared	Reporting biased eta squared or non-comparable partial eta squared
Ordinal or non-normal outcome	Cliff's delta or rank-biserial	Forcing a parametric d onto ranked data
Predictors in a regression	Standardised + unstandardised betas	Standardised betas read as design-invariant

WATCH OUT FOR

Report a confidence interval around every effect size, and build it the right way. For a standardised mean difference the correct interval is noncentral-t based (for example `MBESS::ci.smd` in R, or `esci`), not the point estimate plus or minus 1.96 standard errors, which is too narrow in small samples. For a correlation, construct the interval on the Fisher z scale and back-transform. An effect whose interval runs from trivial to large is a weaker finding than a moderate effect measured tightly, and only the interval, built correctly, shows it.

THE INTERPRETATION

Benchmarks are priors from another literature, not thresholds in yours

Cohen's small, medium, and large were offered as last-resort conventions for a field with no prior estimates. Treating them as decision thresholds is the most common over-read a reviewer flags.

CONVENTIONAL LABELS, AND WHY THEY MISLEAD

Estimator	Often called small	Often called large	Why the label can mislead
Cohen's d	0.2	0.8	Field medians differ; in many areas 0.8 is implausibly large and signals bias
Correlation r	0.1	0.5	Empirical individual-difference medians sit near .1/.2/.3, not .1/.3/.5
Eta squared	0.01	0.14	Positively biased; omega squared is usually smaller
Odds ratio	Near 1	Far from 1	Magnitude depends on baseline risk; not comparable across base rates

PRINCIPLE

Anchor the size to a quantity in the world, not a textbook row. Translate it: a standardised difference into the probability that a random treated case exceeds a random control (the common-language effect size, or probability of superiority); a risk difference into a number needed to treat; a correlation into the share of variance and the practical change at a meaningful predictor shift. A small effect on a serious outcome from a cheap, safe intervention can matter more than a large effect that costs more than it returns.

BEFORE A REVIEWER DOES

Pre-empt the four objections that cost the most points

These are the comments that turn a minor revision into a major one. Each has a one-line defence you can write into the methods before submission.

1. **"Your effect size does not match your test"**
   Tie each estimate to its test: d or g for a t-test, omega squared for an analysis of variance, an odds ratio with its model for logistic regression, rank-biserial for a Mann-Whitney. Reporting eta squared beside a t-test, or a correlation beside a group contrast, reads as software-driven rather than chosen.

2. **"Why standardise by the pooled SD here?"**
   State the standardiser explicitly. If groups have unequal variances, justify Glass's delta or a robust estimator; if you pooled, show the variances were comparable. Silence on the standardiser invites the reviewer to assume the convenient default.

3. **"Your odds ratio overstates the risk"**
   When the outcome is common, convert the odds ratio to a risk ratio for the discussion, or fit a model that targets the risk ratio directly (log-binomial, or Poisson with robust standard errors), and report the baseline risk so readers can judge the gap.

4. **"This effect is implausibly large for the field"**
   A large standardised effect in a small sample is the signature of selective reporting or an underpowered design. Place your estimate against the field's empirical distribution or a relevant meta-analytic mean, and report the interval, so an honest large effect is distinguishable from a fragile one.

BEFORE YOU REPORT IT

The effect size checklist

Every box ticked means the estimator fits the design, the interval is built correctly, and the size is framed against the world rather than a convention.

- [ ] The estimator matches the data-generating design and the estimand you can defend.
- [ ] You reported Hedges' g, not Cohen's d, when either group fell below about twenty.
- [ ] You named the standardiser (pooled, control, or robust) and justified it.
- [ ] For a common binary outcome you chose risk ratio or risk difference, or converted the odds ratio.
- [ ] For variance explained you reported omega or epsilon squared, not unadjusted eta squared.
- [ ] Every effect size carries a confidence interval built on the correct scale (noncentral-t for d, Fisher z for r).
- [ ] You translated the size into a real-world quantity (number needed to treat, probability of superiority).
- [ ] You placed an unusually large effect against the field distribution and reported its precision.

WHEN THE NUMBERS HAVE TO SURVIVE REVIEW

Have your effect sizes computed, bounded, and defended properly

If you want the right estimators calculated for your data or extracted from your studies, with correctly constructed intervals, a robust check, and an interpretation that pre-empts the reviewer, our PhD methodologists do the analysis with you. You stay the author and understand every number before it goes in.

Get a quote for your analysis

researchgold.org . Written by PhD methodologists . We work in writing, on your schedule.