Effect Size and Why Magnitude Matters

A p-value tells you whether an effect is detectable at your chosen threshold. It does not tell you how big that effect is. Effect size is the separate, and often more useful, question of magnitude.

StatReason

Effect Size and Why Magnitude Matters

Four different things people call "effect size"

A raw difference is the plainest form: Group A scored 4 points higher than Group B, in the outcome's original units. A standardized difference, such as Cohen's d or Hedges' g, divides that raw difference by a measure of spread, making it comparable across studies that use different scales. An association measure, such as a correlation coefficient or Cramer's V, describes how strongly two variables move together rather than how far apart two group means sit. A practical importance threshold is a separate, context-specific judgment about how large an effect needs to be before it matters for a real decision, and no statistical formula can set that threshold for you.

Worked example

Two independent groups differ by 4 points on a 100-point test, with a pooled standard deviation of 8 points. The raw difference is 4 points. The standardized difference (Cohen's d) is 4 divided by 8, or 0.5, conventionally described as a medium effect. Whether a 4-point difference on this specific test matters in practice is a separate question the standardized number cannot answer on its own; it depends on what the 100-point scale actually measures and what change would matter to the people using it.

Why standardizing helps, and where it can mislead

Standardized effect sizes let you compare results across studies that measured the same underlying concept on different scales. But standardizing also hides the original units, which can make a trivial real-world difference look impressively sized, or a large real-world difference look small, purely because of how variable the underlying measure happens to be. Always report the raw difference alongside any standardized version.

Convention labels are not universal laws

Small, medium, and large labels attached to standardized effect sizes (Cohen's d around .2, .5, and .8, for instance) are field conventions, not fixed statistical facts. A "small" effect in one field can be enormously consequential, and a "large" effect in another can be practically irrelevant. Every StatReason effect-size result states its formula and lets you judge practical importance in your own context rather than asserting a universal verdict.

Association measures answer yet another question

Cramer's V and similar measures summarize the strength of association in a table of categories; they answer "how related are these two variables," not "how far apart are two group means." Cliff's delta, by contrast, describes ordinal dominance: how often one group's values exceed another's. Each of these measures is a distinct, non-interchangeable way of describing magnitude, matched to a specific kind of data.

What effect size does not mean

A large effect size does not establish causation, and a small one does not prove an effect is unimportant; a tiny per-unit effect can matter enormously at scale. Effect size and statistical significance answer different questions entirely, one of the reasons a result can be statistically significant but practically trivial, or practically important but statistically inconclusive with a small sample. See Statistical Significance vs Practical Significance for that specific contrast.

Source

This page follows the effect-size conventions in the Cochrane Handbook and the shared statistical reasoning contract every StatReason engine is built against.

Limitations

This page does not recommend a universal minimum effect size to consider important; that judgment depends on your field and your specific decision, not on a formula.

Next action: check a raw and standardized difference with the **Independent t-Test** or **Eta Squared Calculator**. Check an association strength with the **Cramer's V Calculator** or **Cliff's Delta Calculator**.