RESEARCH · LITERATURE AND EVIDENCE

Research Synthesis When Findings Conflict

Why studies on the same question often disagree, why counting votes misleads, and how to investigate conflicting findings through heterogeneity analysis, bias checks, context and qualitative synthesis methods.

ModQu research article · 1,681 words · 19 references (APA 7th edition)

Anyone who reviews research on a real question soon finds that studies disagree. One survey finds that price matters most to customers; another finds that service quality does. One trial finds that a programme works; another finds no effect. Conflicting findings are not a sign that research has failed. They are information, and handled well they can show where an effect holds, where it does not and why. This article explains why studies conflict, which common responses mislead, and how to synthesise evidence when findings point in different directions.

Why studies disagree

There are four broad reasons why studies on the same question reach different conclusions.

Chance. Every study based on a sample is subject to sampling error. Small studies in particular can produce estimates that differ widely from the true value, so some disagreement is expected even when every study is sound and the underlying effect is the same.

Real differences. Studies differ in their populations, settings, interventions, measures and time periods. A loyalty scheme may work in one market and not another; a training programme may help new staff but not experienced ones. When the true effect varies between contexts, studies should disagree.

Bias. Studies differ in quality. Poor randomisation, missing data, unreliable measures and selective reporting can push results in one direction. Some disagreement reflects which studies were done well.

Selective publication. Studies with striking or statistically significant results are more likely to be published and cited, so the visible literature may be a biased sample of the research actually done.

The figure below summarises these sources of conflict and the questions each one prompts.

Why studies disagree

Four sources of conflicting findings and the question each one prompts

Source	What it looks like	Question to ask
Chance	Small studies scatter widely	Do the confidence intervals overlap?
Real differences	Effects vary by setting or group	Which study features predict the effect?
Bias	Weaker studies show bigger effects	Do the best studies agree?
Selective publication	Small studies show large effects	Is the funnel plot asymmetric?

Drawing on Thompson (1994) and Borenstein et al. (2009) ModQu.com

Figure 1. Four reasons studies on the same question disagree, and the question each one prompts. Table by ModQu, drawing on Thompson (1994) and Borenstein et al. (2009).

Why vote counting misleads

The simplest response to conflicting studies is to count them: if seven studies found a significant effect and five did not, the effect is judged to exist, or the evidence to be "mixed". This is known as **vote counting**, and methodologists have long warned against it (Borenstein et al., 2009; Hunter & Schmidt, 2004).

Vote counting ignores the size of each study and the size of each effect. A non-significant result from a small study is often consistent with a substantial effect; it simply lacks the precision to rule out zero. If many small studies each have low statistical power, most will fail to reach significance even when a real effect exists, and a vote count will conclude that there is no effect. Conversely, a large study may find a statistically significant effect that is too small to matter in practice. Combining the estimates themselves, weighted by their precision, gives a far better answer than counting outcomes of significance tests.

Consider an illustrative example. Suppose ten small studies each test whether a new onboarding email raises customer retention, and each finds a modest improvement of about three percentage points. Because each study is small, only three of the ten reach statistical significance. A vote count would report "mixed evidence", with seven studies finding no effect. Yet every study pointed in the same direction, and pooling them would give a precise estimate of a consistent, worthwhile benefit. The apparent conflict lies in the significance tests, not in the data.

Measuring how much studies disagree

In a meta-analysis, the extent of disagreement is called **heterogeneity**. Higgins and Thompson (2002) proposed the **I^2 statistic**, which estimates the share of the variation in effect estimates that reflects real differences between studies rather than chance. Values near zero suggest that studies are broadly consistent; higher values suggest substantial real variation. The Cochrane handbook cautions that thresholds for interpreting I^2 are only rough guides, and that the size and direction of the effects matter as much as the statistic itself (Higgins et al., 2019).

When studies are expected to differ, a random-effects model is usually more appropriate than a fixed-effect model. It treats the true effect as varying between studies and estimates the average of that distribution (Borenstein et al., 2009). But an average of very different effects can be misleading if the variation itself is the important finding.

Explaining the disagreement

Thompson (1994) argued that heterogeneity should be investigated rather than merely measured, because the reasons studies differ often hold the most useful lessons. Two techniques are common.

Subgroup analysis compares the pooled effect in groups of studies that differ on a characteristic, such as the type of customer, the country or the length of follow-up.

Meta-regression relates the size of the effect in each study to one or more study characteristics. Thompson and Higgins (2002) warned that meta-regression is observational even when the included studies are trials: studies differ in many ways at once, so an association between a characteristic and effect size may reflect something else. They advised specifying the characteristics to be examined in advance, limiting their number and treating the results as hypotheses rather than conclusions.

Time can also explain conflict. Lau et al. (1992) introduced **cumulative meta-analysis**, adding trials one at a time in order of publication. They showed that the benefit of a treatment for heart attack had been clear from the pooled evidence many years before textbooks and expert reviews recommended it, while some treatments that experts recommended were not supported by the accumulating trial evidence.

Checking for bias

Before concluding that conflicting results reflect real differences, reviewers should ask whether some results are less trustworthy than others.

Study quality. Separating studies at high and low risk of bias, and checking whether they agree, is a basic step. If the better studies show smaller effects, the larger effects in weaker studies may be artefacts.

Publication bias. Rosenthal (1979) described the **file drawer problem**: studies with null results may remain unpublished, so the published literature overstates effects. Franco et al. (2014) confirmed this in the social sciences, finding that most null results from a set of funded experiments were never written up. Egger et al. (1997) proposed a test based on the **funnel plot**, which looks for an asymmetric pattern in which small studies report larger effects than large ones, a pattern consistent with selective publication, although it can also have other causes.

Early and striking results. Ioannidis (2005) followed highly cited clinical studies and found that a substantial share were later contradicted or found to have overstated their effects, particularly when the original studies were small or non-randomised. The Open Science Collaboration (2015) attempted to replicate 100 psychology studies and found that replication effects were, on average, about half the size of the originals. When a striking early study conflicts with later, larger ones, the later evidence usually deserves more weight.

When context explains conflict

Sometimes studies disagree because an intervention works through mechanisms that operate only in certain conditions. Pawson et al. (2005) developed **realist review** to address this. Rather than asking whether an intervention works on average, a realist review asks what works, for whom, in what circumstances and why. Conflicting findings are treated as clues: a programme that succeeded in some places and failed in others may have triggered a mechanism in one context that was absent or blocked in another. Cartwright and Hardie (2012) made a similar argument for policy, holding that evidence from one setting supports predictions in another only when the same supporting factors are present.

This approach is valuable in business research, where the same marketing campaign, pricing change or management practice may perform very differently across markets, customer segments and organisations. A synthesis that identifies the conditions for success is more useful than one that reports an average effect.

Conflict in qualitative and mixed evidence

Qualitative studies can also disagree. Noblit and Hare (1988), in developing meta-ethnography, described three kinds of synthesis: **reciprocal translation**, when studies describe similar concepts in different terms; **refutational synthesis**, when studies contradict one another; and **lines-of-argument synthesis**, which builds a fuller interpretation from studies that each show part of a picture. Refutational synthesis treats contradiction as something to explain, by asking how differences in participants, settings or the researchers' perspectives might account for it.

Thomas and Harden (2008) developed **thematic synthesis**, which codes the findings of qualitative studies, develops descriptive themes and then builds analytical themes that go beyond the original studies. Their method has been used to explain why interventions that look promising in trials do not always succeed in practice, by setting trial results alongside qualitative evidence on what people want and need.

Dixon-Woods et al. (2005) reviewed methods for combining qualitative and quantitative evidence and noted that each involves choices about whether the aim is to aggregate findings or to interpret them. Pope et al. (2007) described a range of approaches for synthesising mixed evidence in health research, from narrative and thematic methods to Bayesian approaches, and stressed that the choice should follow from the question. In mixed reviews, an apparent conflict between quantitative and qualitative findings may show that a measured outcome missed something that mattered to participants.

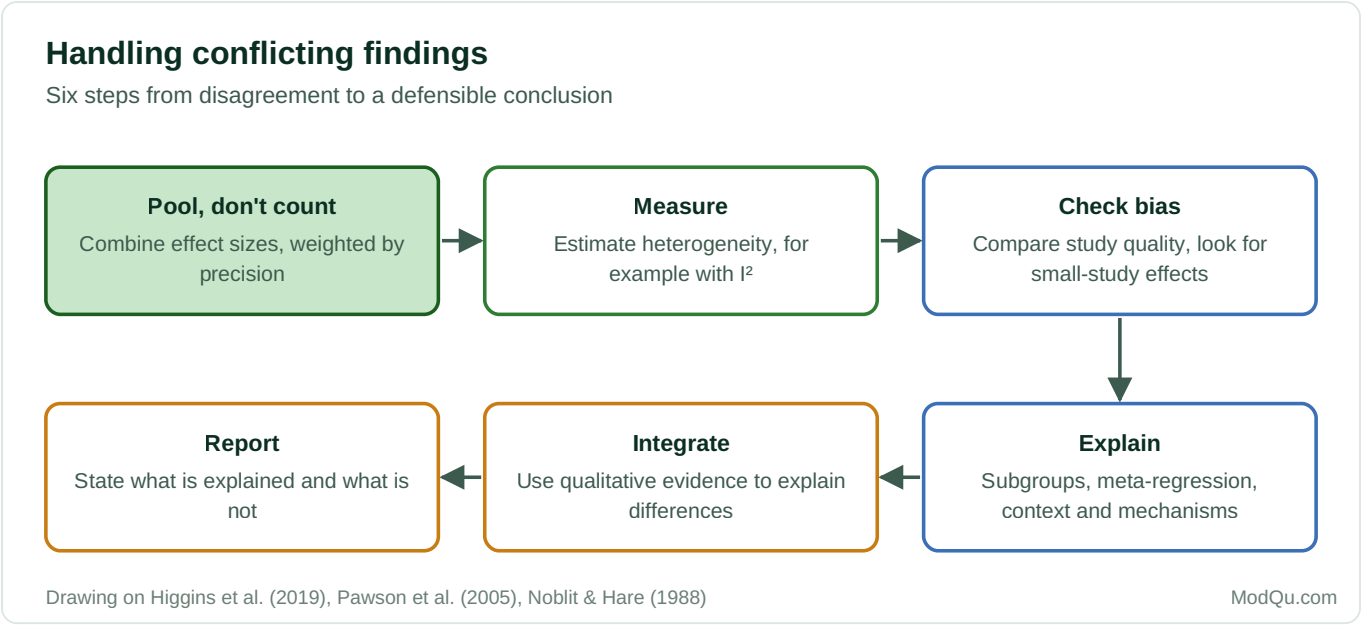


Figure 2. Six steps for synthesising conflicting findings. Diagram by ModQu, drawing on Higgins et al. (2019), Pawson et al. (2005) and Noblit and Hare (1988).

Reporting conflicting evidence

Clear reporting matters as much as analysis. The GRADE approach treats unexplained inconsistency as one reason to lower the certainty of a body of evidence (Guyatt et al., 2008). A good synthesis states how much studies disagree, what was done to explain the disagreement, what explanations were supported and what remains unexplained. Readers should be able to see whether a conclusion holds across the evidence or depends on a few studies, and whether it applies to their own context.

Conclusion

Conflicting findings are a normal part of research. They arise from chance, from real differences between settings, from bias and from selective publication. Counting studies on each side hides these causes. A careful synthesis measures the disagreement, investigates it through subgroups, bias checks and attention to context, and reports what it found plainly. Handled in this way, disagreement between studies often teaches more than agreement would.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: A review of possible methods. *Journal of Health Services Research & Policy, 10*(1), 45–53. https://doi.org/10.1177/135581960501000110

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ, 336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley. https://doi.org/10.1002/9781119536604

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). SAGE. https://doi.org/10.4135/9781412985031

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA, 294*(2), 218–228. https://doi.org/10.1001/jama.294.2.218

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine, 327*(4), 248–254. https://doi.org/10.1056/NEJM199207233270406

Noblit, G. W., & Hare, R. D. (1988). *Meta-ethnography: Synthesizing qualitative studies*. SAGE. https://doi.org/10.4135/9781412985000

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review: A new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy, 10*(Suppl. 1), 21–34. https://doi.org/10.1258/1355819054308530

Pope, C., Mays, N., & Popay, J. (2007). *Synthesizing qualitative and quantitative health evidence: A guide to methods*. Open University Press.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology, 8*, Article 45. https://doi.org/10.1186/1471-2288-8-45

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *BMJ, 309*(6965), 1351–1355. https://doi.org/10.1136/bmj.309.6965.1351

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*(11), 1559–1573. https://doi.org/10.1002/sim.1187