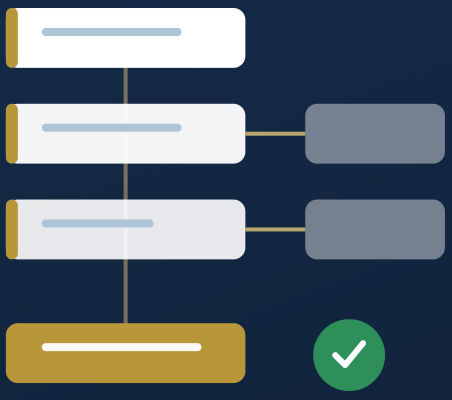
Research Gold

RESEARCH GOLD FIELD GUIDE

Rating Certainty of Evidence with GRADE

How to judge how much confidence your findings deserve, the step that turns a synthesis into a usable conclusion, explained without jargon. A working guide to GRADE from PhD methodologists.



Written by PhD methodologists researchgold.org

START HERE

A finding is only as useful as the confidence behind it

A pooled estimate or a synthesised pattern is not the end of a review. The reader needs to know how much to trust it. GRADE is the structured method for rating the certainty of evidence, from high to very low, so your conclusions carry an honest signal of their own strength. It is increasingly expected in any review meant to inform a decision.

PRINCIPLE

Certainty is rated per outcome, not for the review as a whole. A review can have high certainty for one outcome and low for another, because the evidence differs. Rate each important outcome separately, and state the certainty alongside the result so the reader weighs it correctly.

How a rating is built

1. **Start from the study design**
   Evidence from randomised trials starts at high certainty. Evidence from observational studies starts at low. This is the starting point, not the verdict.

2. **Rate down for problems**
   Five concerns can lower the rating: risk of bias, inconsistency between studies, indirectness, imprecision, and publication bias. Each serious concern moves the rating down a level.

3. **Rate up where warranted**
   Observational evidence can be rated up for a very large effect, a clear dose-response relationship, or when plausible confounding would only have reduced the observed effect.

4. **Land on a certainty level**
   The result is one of four levels: high, moderate, low, or very low. Each carries a plain-language meaning about how likely the true effect is close to the estimate.

THE FIVE CONCERNS

What can lower your certainty

Work through each of the five domains for every important outcome. A serious or very serious problem in any one lowers the rating.

THE FIVE REASONS TO RATE DOWN

Concern	The question it asks	When to rate down
Risk of bias	Were the studies well conducted?	When studies driving the result are at high risk
Inconsistency	Do the studies agree?	When effects vary widely with no good explanation
Indirectness	Do the studies match your question?	When population, intervention, or outcome differ
Imprecision	Is the estimate precise enough?	When the confidence interval spans different decisions
Publication bias	Are missing studies likely?	When small, positive studies dominate the evidence

WATCH OUT FOR

Do not rate down twice for the same underlying problem. If high risk of bias also explains why studies disagree, count it once, under the domain that fits best. Double-counting a single weakness drives the certainty artificially low and misrepresents the evidence.

PRINCIPLE

Imprecision is the domain most often misapplied. It is not simply "does the confidence interval cross the null". GRADE judges imprecision two ways: whether the interval crosses a clinically important threshold (a minimally important difference) such that it spans both a worthwhile and a trivial effect, and whether the total sample meets the optimal information size, the number of participants a single adequately powered trial would require. A result can exclude the null yet still be rated down for imprecision if the evidence base is too small to meet that information size, or rated as precise despite touching the null if the whole interval lies within trivial effects. Rate against the decision threshold, not against zero.

THE OUTPUT

Present certainty so a reader can act on it

The result of GRADE is usually a summary of findings table, with a certainty rating and a plain-language statement for each outcome.

Summary of findings row (one per outcome)

```
Outcome:          [name the outcome]
Effect estimate:  [pooled or synthesised result, with interval]
Participants:     [number, across number of studies]
Certainty:        [High / Moderate / Low / Very low]
Reason(s) down:   [which of the five concerns, and how serious]
Plain statement:  [for example: the intervention probably reduces
                  [outcome] (moderate certainty)]
```

PRINCIPLE

Translate each certainty level into the words it implies. High certainty supports "the intervention reduces"; moderate supports "probably reduces"; low supports "may reduce"; very low means the effect is uncertain. Matching your wording to the certainty stops a review from claiming more than its evidence allows.

BEFORE YOU STATE YOUR CONCLUSIONS

The certainty rating checklist

If every box is ticked, your conclusions will carry an honest, structured signal of how far they can be trusted.

- [x] You rated certainty separately for each important outcome.
- [x] You set the correct starting level from the study design.
- [x] You considered all five reasons to rate down for each outcome.
- [x] You judged imprecision against a clinically important threshold and the optimal information size, not just whether the interval crosses the null.
- [x] You rated up only where the recognised criteria genuinely apply.
- [x] You avoided rating down twice for the same underlying problem.
- [x] You presented the results in a summary of findings table.
- [x] Your wording for each conclusion matches its certainty level.

WHEN THE CONCLUSIONS HAVE TO BE TRUSTED

Have the certainty of your evidence rated properly

If you want each outcome rated with GRADE, presented in a clear summary of findings table, and your conclusions worded to match their certainty, our PhD methodologists do it with you as part of a synthesis that holds up in peer review. You stay the author and make every judgement call.

Get a quote for your review

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