SCREENING

Agreement between screeners: percent, Cohen’s kappa and PABAK

For two people who screened the same records, and the supervisor who reads their figures.

By the end you will have your agreement worked out three ways, and a methods sentence that says how.

Checked on 29 Sep 2026 · Version 1.0 · CC BY 4.0

When two people screen the same records, a supervisor, an examiner or a journal may ask how well they agreed. The answer is a small table and a few numbers worked out from it.

This guide builds the table, works out percent agreement, Cohen’s kappa and PABAK, and shows why kappa can look low when you agreed on nearly every record. It ends with a sentence for your methods section.

Before you start

- Both of you screened the same records, each without seeing the other’s decisions.
- You have both sets of decisions, matched record by record.
- You know which stage you are measuring. Titles and abstracts, and full texts, each get their own figures.
- You have chosen how to count Maybe (see “Count Maybe one way, and say which”).

Check what you are asked for

PRISMA 2020 item 8 asks you to specify “how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation

tools used in the process" (Page et al. 2021). Its explanation and elaboration adds how disagreements were resolved. Neither asks for an agreement statistic.

Table 1. What the main sources say about agreement (checked 29 Sep 2026)

Source	What it says
PRISMA 2020, item 8	Report how many people screened, whether they worked independently, and any automation. No statistic.
Cochrane Handbook, chapter 4	No agreement statistic or threshold for screening.
Cochrane Handbook, 5.5.5 (collecting data)	Agreement on coded data items can be measured with kappa, "although this is not routinely done in Cochrane reviews".
Cochrane Handbook, 7.3.2 (risk of bias)	Advises against kappa for describing agreement; exploring and resolving the reasons for disagreement matters more.
Cochrane Rapid Reviews Methods Group, tutorial (2026)	Dual screening of a proportion of records, then single screening if agreement is high enough. Its example of high enough is a Cohen's kappa of 0.80 or more.
JBI Manual, scoping reviews, 10.2.7.1	A pilot on 25 random titles and abstracts; screening starts at 75% agreement or more. The measure is percent agreement.

So kappa is a figure you choose to give, or one your programme or journal asks for. Their instructions may ask for more than the standards do, so read them first.

Build the 2×2 table

Put one person's decisions in the rows and the other's in the columns, and count the records in each cell. Count only records both of you decided.

Table 2. The 2×2 table

	B includes	**B excludes**
A includes	a	b
A excludes	c	d

The cells a and d are the records you agreed on. The cells b and c are the two kinds of difference. The total, n = a + b + c + d, is the number of records both of you screened.

Work out the three figures

Four formulas do all of it.

- **Observed agreement:** $p_0 = (a + d) / n$. Multiplied by 100, this is percent agreement.
- **Agreement expected by chance:** $p_e = [(a + b)(a + c) + (c + d)(b + d)] / n^2$. This is how often you would agree if each of you kept your own rate of including but chose at random.

- **Cohen's kappa:** $\kappa = (p_0 - p_e) / (1 - p_e)$. It is the share of the agreement possible beyond chance that you reached.
- **PABAK:** $2p_0 - 1$. See "Add a figure that allows for prevalence".

A worked example

Two people screen 2,000 records. Both include 90 and both exclude 1,810. A includes 40 that B excludes, and B includes 60 that A excludes.

Table 3. The worked example (illustrative numbers)

	B includes	B excludes	Total
A includes	90	40	130
A excludes	60	1,810	1,870
Total	150	1,850	2,000

- $p_0 = (90 + 1{,}810) / 2{,}000 = 0.95$, so percent agreement is 95%.
- $p_e = (130 \times 150 + 1{,}870 \times 1{,}850) / 2{,}000^2 = 3{,}479{,}000 / 4{,}000{,}000 = 0.86975$.
- $\kappa = (0.95 - 0.86975) / (1 - 0.86975) = 0.08025 / 0.13025 = 0.616$, or 0.62 to two places.
- PABAK $= 2 \times 0.95 - 1 = 0.90$.

Round only at the end.

Spreadsheet formulas

Type a in B2, b in C2, c in B3 and d in C3, then these formulas. They are the same in Excel and Google Sheets.

Excel · the same in Google Sheets · a in B2, b in C2, c in B3, d in C3

```
B5 (n)       =SUM(B2:C3)
B6 (p0)      =(B2+C3)/B5
B7 (pe)      =((B2+C2)*(B2+B3)+(B3+C3)*(C2+C3))/B5^2
B8 (kappa)   =(B6-B7)/(1-B7)
B9 (PABAK)   =2*B6-1
```

Copy the formulas from the web page, where each copy button gives the exact text: evensift.com/guides/cohens-kappa-screening/

With the worked example, B8 comes to 0.616 and B9 to 0.9.

Agreement from four counts

Each record both reviewers screened goes in one of four cells.

	Reviewer 2 included	Reviewer 2 excluded
Reviewer 1 included	a	b
Reviewer 1 excluded	c	d

n, records both screened	$a + b + c + d$
p_0, observed agreement	$(a + d) / n$
p_e, chance agreement	$[(a + b)(a + c) + (c + d)(b + d)] / n^2$
Cohen's κ	$(p_0 - p_e) / (1 - p_e)$
PABAK	$2p_0 - 1$

See why kappa runs low when you agree

In the worked example you agreed on 95% of records, and kappa is 0.62. That can read like a verdict on your screening. It is mostly arithmetic. When nearly every record is excluded, two people agree on most of them by chance alone. Here chance accounts for 0.870 of the scale, so kappa is measured within the 0.130 that is left.

Table 4 shows the same differences in two sets of 1,000 records.

Table 4. The same differences, two kappas (illustrative numbers)

Set	a	b	c	d	Included by each	p_0	p_e	κ	PABAK
Typical screen	40	30	30	900	7%	0.94	0.870	0.54	0.88
Balanced set	440	30	30	500	47%	0.94	0.502	0.88	0.88

Both sets have 60 differences and 94% agreement. Only the share of records included changes, and kappa moves from 0.54 to 0.88.

Byrt, Bishop and Carlin (1993) showed that kappa depends on prevalence (how records fall across the categories) and on bias between the two observers, as well as on agreement. They recommended reporting measures of both beside it. Feinstein and Cicchetti (1990) named the effect: high agreement but low kappa. Your 2×2 table shows prevalence and bias at a glance, which is why it belongs in your supplement.

Add a figure that allows for prevalence

PABAK is the prevalence-adjusted bias-adjusted kappa (Byrt, Bishop and Carlin 1993). For two categories it is $2p_0 - 1$ (Chen et al. 2009). It assumes a prevalence of 50% and no bias between the screeners, so chance agreement is 0.5, and kappa's formula becomes $(p_0 - 0.5) / (1 - 0.5) = 2p_0 - 1$.

PABAK is observed agreement moved onto kappa's scale. It adds nothing that p_0 does not hold, but it lets a reader set kappa beside a figure that prevalence has not pulled down. Gwet (2008) proposed another coefficient, AC1, as a more stable alternative to kappa when agreement is high. Report either beside kappa, not instead of it.

Read the number with its table

Published scales attach words to ranges of kappa. They are conventions, not standards. McHugh (2012) argues that the usual ones may be too lenient for health research, and suggests calculating both percent agreement and kappa. Put the table beside the figures, and a reader can judge them for themselves.

Count Maybe one way, and say which

None of the standards above says how to count Maybe. Choose before you count, and name your choice in the methods.

- **Maybe counted as Include.** If your Maybes go forward to full text, they act as includes, and the table stays 2×2.
- **Maybe as a third category.** The table becomes 3×3. The agreed records are the three cells on the diagonal, and p_e adds three products instead of two. The two-category PABAK formula no longer applies.

Whichever you choose, apply it to both screeners and to every record.

Say which records the figure covers

A figure from a pilot describes the pilot. Say whether yours covers a pilot set, a proportion of records, or every record. For a sense of size: the Cochrane Handbook suggests piloting the criteria on about six to eight reports (section 4.6.4), JBI's scoping-review pilot is 25 records, and the Rapid Reviews group suggests "e.g., 20–50 citations".

Talk it through when agreement is low

A low figure is a reason to talk. Go through the differences one at a time, and look for a criterion the two of you read differently. Reword it, give your criteria a new version and date, and pilot again on records neither of you has seen. The Cochrane Handbook says disagreements can usually be settled by discussion, with another person to arbitrate if needed (section 4.6.4).

Check your work

- a + b + c + d equals the number of records both of you screened.
- κ cannot be larger than p_0, and nor can PABAK. If either is, look again at the formulas.
- If both of you included every record, or both excluded every record, p_e is 1 and kappa cannot be worked out. Report percent agreement instead.
- Put the worked example's four counts through your spreadsheet first. It should give 0.616 for κ and 0.9 for PABAK.

Writing it up

A sentence to adapt:

> Two reviewers ([initials]) independently screened all [n] titles and abstracts. Disagreements were resolved by discussion, or by a third reviewer ([initials]) where needed. Agreement was [x]% (Cohen's κ = [k]; PABAK = [p]), with Maybe counted as Include; the 2×2 table is in Supplement [S]. No automation tools were used to exclude records.

With the worked example, the third sentence begins "Agreement was 95% (Cohen's κ = 0.62; PABAK = 0.90)". If you used an automation tool at any point, say how in place of the last sentence: item 8 asks for it.

Reading on

- Screening alone or with a colleague
- How to screen titles and abstracts
- McHugh ML. Interrater reliability: the kappa statistic (2012) (pmc.ncbi.nlm.nih.gov/articles/PMC3900052/)
- PRISMA 2020 explanation and elaboration (2021) (doi.org/10.1136/bmj.n160)

Sources

1. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. Item 8. CC BY 4.0. https://doi.org/10.1136/bmj.n71, checked 29 Sep 2026.
2. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 2021;372:n160. Item 8. https://doi.org/10.1136/bmj.n160, checked 29 Sep 2026.
3. Lefebvre C, Glanville J, Briscoe S, et al. Chapter 4: Searching for and selecting studies [last updated March 2025]. *Cochrane Handbook for Systematic Reviews of Interventions* version 6.5.1, section 4.6.4. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-04, checked 29 Sep 2026.
4. Li T, Higgins JP, Deeks JJ. Chapter 5: Collecting data [last updated October 2019]. *Cochrane Handbook* version 6.5, section 5.5.5. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-05, checked 29 Sep 2026.
5. Boutron I, Page MJ, Higgins JP, et al. Chapter 7: Considering bias and conflicts of interest among the included studies [last updated August 2022]. *Cochrane Handbook* version 6.5, section 7.3.2. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-07, checked 29 Sep 2026.
6. Garritty C, Hamel C, Nussbaumer-Streit B. Rapid reviews: a tutorial. *Cochrane Evid Synth Methods* 2026;4(4):e70092, section 3.4. CC BY 4.0. https://pmc.ncbi.nlm.nih.gov/articles/PMC13387409/, checked 29 Sep 2026.
7. Pollock D, Peters MDJ, Tricco AC, et al. Chapter 10: Scoping reviews (2026), section 10.2.7.1. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, eds. *JBI Manual for Evidence Synthesis*. JBI; 2024. https://doi.org/10.46658/JBIMES-24-09, checked 29 Sep 2026.
8. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol* 1993;46(5):423–9. Abstract. https://pubmed.ncbi.nlm.nih.gov/8501467/, checked 29 Sep 2026.
9. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43(6):543–9. Abstract. https://pubmed.ncbi.nlm.nih.gov/2348207/, checked 29 Sep 2026.
10. Chen G, Faris P, Hemmelgarn B, Walker RL, Quan H. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. *BMC Med Res Methodol* 2009;9:5. https://doi.org/10.1186/1471-2288-9-5, checked 29 Sep 2026.
11. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol* 2008;61(1):29–48. Abstract. https://pubmed.ncbi.nlm.nih.gov/18482474/, checked 29 Sep 2026.
12. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;22(3):276–82. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052/, checked 29 Sep 2026.
13. Microsoft Support. Excel functions (alphabetical). https://support.microsoft.com/en-us/office/excel-functions-alphabetical-b3944572-255d-4efb-bb96-c6d90033e188, checked 29 Sep 2026.
14. Google Docs Editors Help. Google Sheets function list. https://support.google.com/docs/table/25273, checked 29 Sep 2026.

Cite this guide. Evensift. Agreement between screeners: percent, Cohen’s kappa and PABAK. Evensift guides, version 1.0, 29 Sep 2026. https://evensift.com/guides/cohens-kappa-screening/

Version 1.0 · 29 Sep 2026 · What changed: First publication. The latest version is always at https://evensift.com/guides/cohens-kappa-screening/. If something here no longer matches what you see, write to hello@evensift.com. Corrections are made within 7 days, with a dated note at the foot of this page.