Docket

DOCKET / RESEARCH & PRACTICAL GUIDANCE

Cohen's Kappa for Acquisition Screening Reviewer Agreement

> Build a more reviewable acquisition pipeline with Docket. Our deal-origination software and managed research help private equity teams screen companies against a mandate and understand the evidence behind each finding.

Published by Docket · 2026-10-29 · inter rater reliability · cohens kappa · reviewer agreement screening · inter-rater reliability meaning

Original article: https://docket.capital/articles/cohens-kappa-reviewer-agreement-screening



In this article

1. Why reviewer agreement is a process metric
2. Percent agreement and its blind spot
3. Cohen's kappa step by step
4. Reading the number: bands and their limits
5. When kappa misleads: the prevalence problem
6. What low agreement tells you about the rubric
7. Adjudication: what happens to the disagreements
8. Sample size and cadence
9. Sources

Why reviewer agreement is a process metric

When two analysts apply the same screening rubric to the same company, they should reach the same answer. If they do not, the disagreement is information about the rubric, the evidence or the reviewers, and it is cheap to collect before it turns into a missed target or a wasted outreach slot.

Inter-rater reliability is the general name for this property. McHugh defines it as "the extent to which data collectors (raters) assign the same score to the same variable" (McHugh 2012). For an origination team the "variable" is a screening call: pass or fail on a size band, a disqualifier hit, an ownership type, a fit score.

One distinction has to be fixed before any number is computed. **Agreement measures consistency, not accuracy.** Two reviewers can agree on every record and both be wrong, for example because both trusted the same stale vendor field. Agreement statistics tell you whether the rubric is being applied the same way; accuracy needs a separate check against verified evidence, such as the record-level re-review described in the audit sampling worksheet. A team needs both, and this article covers the first.

Percent agreement and its blind spot

The simplest measure is percent agreement: the number of records on which both reviewers gave the same call, divided by the number of records both reviewed. McHugh describes it as "calculated as the number of agreement scores divided by the total number of scores" (McHugh 2012).

Percent agreement is easy to compute and easy to explain. Its weakness is that some agreement happens by chance. In 1960 Jacob Cohen "critiqued use of percent agreement due to its inability to account for chance agreement" and introduced kappa to correct for it (McHugh 2012; original paper: Cohen 1960).

The chance problem is sharp in acquisition screening because most screens are lopsided. If 90% of a raw list fails a revenue floor, two reviewers who each fail 90% of records at random would still agree on a large share of them. A high percent agreement on a list like that says little about whether the reviewers are reading the rubric the same way.

Cohen's kappa step by step

Cohen's kappa compares observed agreement with the agreement expected if the two reviewers' calls were independent, given how often each reviewer says "pass" and "fail". The formula, as given by both McHugh and Hallgren, is:

$\kappa = (Pr(a) - Pr(e)) / (1 - Pr(e))$

where Pr(a) is observed agreement and Pr(e) is chance agreement (McHugh 2012; Hallgren 2012). Chance agreement is computed from the row and column totals (the marginals) of the two reviewers' cross-tabulation (McHugh 2012).

Kappa can range from −1 to +1, "where 0 represents the amount of agreement that can be expected from random chance, and 1 represents perfect agreement between the raters" (McHugh 2012). Hallgren adds that −1 indicates "perfect" disagreement (Hallgren 2012).

Illustrative example (fictional): a 100-record pass/fail screen

A fictional search fund screens a list of HVAC services companies against a written rubric. Two analysts, A and B, independently review the same 100 records and mark each "pass" (advance to research) or "fail". The cross-tabulation is:

	B: pass	B: fail	Row total (A)
A: pass	20	10	30
A: fail	5	65	70
Column total (B)	25	75	100

Step 1. Observed agreement. The reviewers agree on the diagonal: 20 + 65 = 85 records. Pr(a) = 85 / 100 = 0.85.

Step 2. Expected (chance) agreement. A passes 30% of records and B passes 25%. If their calls were independent, both would say "pass" on 0.30 × 0.25 = 0.075 of records and both would say "fail" on 0.70 × 0.75 = 0.525. Pr(e) = 0.075 + 0.525 = 0.60. In McHugh's marginal form: ((30 × 25) / 100 + (70 × 75) / 100) / 100 = (7.5 + 52.5) / 100 = 0.60.

Step 3. Kappa. κ = (0.85 − 0.60) / (1 − 0.60) = 0.25 / 0.40 = **0.625**.

Percent agreement says 85%. Kappa says the reviewers achieved 62.5% of the agreement that was available beyond chance. Both numbers are correct; they answer different questions.

Step 4. Look at the disagreements, not just the number. The 15 off-diagonal records are not symmetric. A passed 10 records that B failed; B passed only 5 that A failed. A is more permissive than B. That pattern is a lead, not a verdict: it may mean A treats a missing revenue figure as "unknown, advance" while B treats it as "fail", which is a rubric problem, not a reviewer problem.

Reading the number: bands and their limits

Teams want a threshold. The published ones are conventions, and they are contested.

McHugh reports Cohen's suggested interpretation: values ≤ 0 as no agreement, 0.01–0.20 none to slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial and 0.81–1.00 almost perfect (McHugh 2012). Hallgren reports the widely cited Landis and Koch guidelines in nearly the same form: 0.0–0.2 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial and 0.81–1.0 almost perfect or perfect agreement, and notes that the use of such qualitative cutoffs is debated (Hallgren 2012; original: Landis and Koch 1977).

Kappa	Label (Landis and Koch, via Hallgren)	Practical reading for a screening team
0.00–0.20	Slight	The rubric is not functioning as a shared standard
0.21–0.40	Fair	Expect many coin-flip calls; rewrite criteria before scaling
0.41–0.60	Moderate	Below McHugh's adequacy line; diagnose before relying on single review

Kappa	Label (Landis and Koch, via Hallgren)	Practical reading for a screening team
0.61–0.80	Substantial	Usable, with double review on high-stakes calls
0.81–1.00	Almost perfect	Monitor on a sample; watch for drift

The third column is practice guidance from this article, not a published standard.

McHugh argues the conventional labels are too lenient for work where errors matter, because they allow a score as low as 0.41 to be called "moderate", and concludes that "any kappa below 0.60 indicates inadequate agreement among the raters and little confidence should be placed in the study results" (McHugh 2012). She also notes that "many texts recommend 80% agreement as the minimum acceptable interrater agreement" for percent agreement (McHugh 2012).

By either reading, the fictional 0.625 above is just over the line. It would justify continuing with the rubric, but not trusting a single reviewer on records that decide whether a company is contacted.

Put an interval around the estimate

Kappa from 100 records is an estimate. McHugh gives an approximate standard error, $SE = \sqrt{(Pr(a)(1 - Pr(a)) / (n(1 - Pr(e))^2)}$, and a 95% interval of $\kappa \pm 1.96 \times SE$ (McHugh 2012). For the example: $SE = \sqrt{(0.85 \times 0.15 / (100 \times 0.40^2))} = \sqrt{(0.1275 / 16)} \approx 0.089$, so the interval is roughly 0.625 ± 0.175, or **0.45 to 0.80**. That range spans "moderate" to "substantial". McHugh notes that small samples produce wide intervals and suggests, as a general heuristic, not using fewer than 30 comparisons (McHugh 2012). Report the interval with the point estimate.

When kappa misleads: the prevalence problem

Kappa's correction for chance depends on the marginals, and that creates a known failure mode. Hallgren states that prevalence problems "typically [cause] kappa estimates to be unrepresentatively low", that bias problems (reviewers with different overall pass rates) typically make kappa unrepresentatively high, and that "no single kappa variant corrects for both", so multiple variants may need to be reported (Hallgren 2012). The classic paper on the issue, Feinstein and Cicchetti's "High agreement but low kappa: I. The problems of two paradoxes", states the first paradox in its abstract: in a two-observer fourfold table, "a high value of p0 can be drastically lowered by a substantial imbalance in the table's marginal totals", p0 being observed agreement (Feinstein and Cicchetti 1990).

The arithmetic shows why this matters for screening. Take two fictional 100-record reviews with identical percent agreement (92%):

Fictional scenario	Both pass	A pass / B fail	A fail / B pass	Both fail	Percent agreement	Pr(e)	Kappa
Balanced list (half pass)	46	4	4	46	92%	0.50	0.84
Skewed list (6% pass)	2	4	4	90	92%	0.887	0.29

The reviewers disagree on exactly 8 records in both cases. On the skewed list, chance agreement is already 88.7% because both reviewers fail almost everything, so the same 8 disagreements leave little room above chance and kappa falls to 0.29. The skewed case is also fragile: with only 6 passes each, McHugh's standard error formula gives a 95% interval from about −0.18 to 0.76.

This is the normal shape of a top-of-funnel screen. A raw list of several thousand companies may yield a small share of passes, and kappa on that list will look poor even when reviewers agree on nearly every record. Two practical responses:

1. **Always report the full 2×2 table, percent agreement and each reviewer's pass rate alongside kappa.** McHugh's own advice is to "calculate both percent agreement and kappa" (McHugh 2012).
2. **Measure agreement on the decisions that matter.** Agreement on the small set of records either reviewer passed tells you more about the pass/fail boundary than agreement on thousands of obvious fails.

What low agreement tells you about the rubric

A low kappa is a symptom. Before retraining anyone, read the disagreements and classify each one. In practice, most fall into four groups:

Disagreement type	What it looks like	Fix
Ambiguous criterion	"Recurring revenue majority" read as contracts by one reviewer and repeat customers by the other	Rewrite the criterion with a definition and an example of each side of the line
Missing-evidence rule	One reviewer fails a record with no revenue evidence, the other advances it as "unknown"	Write an explicit rule for unknowns; keep "unknown" as its own outcome
Evidence divergence	Reviewers looked at different sources (website vs registry)	Specify the source order per field; see the claim ledger for conflicting sources
Definition drift	Agreement was high in month one and fell later	Re-test on a fresh sample after every rubric change

Disqualifiers deserve their own measurement. A disqualifier that one reviewer applies and another misses changes outcomes directly, so compute agreement separately for each knockout question rather than only for the overall pass/fail call. Asymmetric disagreement, as in the fictional example (10 versus 5), usually points to a missing-evidence rule rather than to one careless reviewer.

After changing the wording, re-test. A fictional re-run of the same exercise on a fresh 100 records, after the "unknown revenue" rule was written down, might produce 24 / 3 / 2 / 71: percent agreement 95%, Pr(e) about 0.61, kappa about 0.87. The point of the illustration is the loop, not the number: measure, read the disagreements, fix the text, measure again.

Adjudication: what happens to the disagreements

Agreement statistics are computed on the independent first calls. Production decisions still need a single answer per record. A workable process:

1. **Blind first review.** Both reviewers record their call and a one-line reason without seeing each other's work. The statistic is only valid if the calls were independent.
2. **Log both calls before resolving.** Keep the original calls, reasons and evidence references. Overwriting them destroys the data the rubric fix depends on.
3. **Resolve by rule, then by third reviewer.** Some disagreements resolve mechanically (one reviewer missed a source). The rest go to a third reviewer or the deal lead, who records the final call and which criterion caused the split.
4. **Feed the log back.** Count disagreements per criterion each cycle. A criterion that produces repeated splits is a drafting problem.

Field	Example (fictional)
Record ID	T-0417
Reviewer A call / reason	Pass: "revenue unknown, all other criteria met"
Reviewer B call / reason	Fail: "no revenue evidence"
Criterion causing split	Revenue floor (unknown-value rule)
Resolution method	Rule added: unknown revenue = hold, not pass
Final call	Hold
Rubric change logged	Yes, v1.3

Sample size and cadence

Double review costs time, so most teams double-review a sample rather than every record. Hallgren notes that when ratings are costly, "selecting a subset of subjects for IRR analysis may be more practical", and that the subset's reliability may be used to generalize to the full sample (Hallgren 2012).

Three practical rules follow from the sources above, without inventing a Docket threshold:

- **Use enough records for a usable interval.** McHugh's heuristic floor is 30 comparisons (McHugh 2012); with lopsided pass rates you will need more, or a sample enriched with records near the boundary.
- **Re-measure after every rubric change** and when new reviewers join.
- **Draw the sample at random** from the records actually screened, and record how it was drawn. The audit sampling worksheet covers sample selection for accuracy audits; the same discipline applies here.

Agreement worksheet

Item	Value
Rubric version	
Criterion measured (overall or single disqualifier)	
Records double-reviewed (n) and how drawn	
2×2 counts (both pass / A only / B only / both fail)	
Reviewer A pass rate; reviewer B pass rate	
Percent agreement	
Pr(e)	
Kappa and 95% interval	
Top criterion causing splits	
Action taken and re-test date	

Docket runs a separate Audit agent that re-reads each researched record with fetch tools denied, which addresses accuracy rather than reviewer agreement; teams with two human reviewers can apply the worksheet above as written.

Sources

1. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica, 2012;22(3):276-282 (PubMed Central). Accessed 2026-10-28.
2. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in Quantitative Methods for Psychology, 2012;8(1):23-34 (PubMed Central). Accessed 2026-10-28.
3. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 1960;20(1):37-46 (closed access; cited as the original paper, no finding taken from it directly). Accessed 2026-10-28.
4. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics, 1977;33(1):159-174 (abstract read on PubMed, PMID 843571; interpretation bands cited via Hallgren). Accessed 2026-10-28.
5. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 1990;43(6):543-549 (abstract read on PubMed, PMID 2348207). Accessed 2026-10-28.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052

2. https://doi.org/10.1177/001316446002000104
3. https://pmc.ncbi.nlm.nih.gov/articles/PMC3402032
4. https://doi.org/10.2307/2529310
5. https://doi.org/10.1016/0895-4356%2890%2990158-L

THE PUBLISHER

About Docket

Docket provides deal-origination research software and managed research for private equity firms. We help investment teams investigate acquisition targets using structured screening criteria, retained sources and reviewable company evidence. Teams can work through a self-serve platform or use managed research, depending on how they want research delivered.

Research against a defined mandate

A useful target list needs more than company names. Docket focuses on the questions that determine whether a company fits an investment mandate, including the evidence needed to support or qualify each answer. Our research approach makes the connection between screening criteria, source material and conclusions visible to the team reviewing the work.

Triage, Scout and Audit

Docket's three named research agents perform complementary tasks. Triage resolves company identity and screens fit. Scout collects sourced answers against the mandate. Audit checks retained evidence, addresses contradictions and leaves unsupported answers visibly unresolved. This structure helps reviewers distinguish established findings from missing information and questions requiring further investigation.

Evidence that supports investment-team judgment

Our research library covers market mapping, screening criteria, private-company data, succession and ownership, source evaluation and evidence standards. These resources explain the methods and limitations behind origination research. Findings support a team's judgment; they do not establish that a company is for sale or guarantee a transaction or investment outcome.

Work with Docket

Visit Docket to explore the platform, managed research and the current contact path. Read about screening criteria, evidence standards, AI research agents and market mapping.

Public examples are illustrative unless explicitly identified otherwise. Research preparation and authorized outreach are separate activities; confidential target lists and customer outcomes should never be inferred from an educational example.

Website: https://docket.capital

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.