Cold Email Deliverability Scorecard

Ten dimensions of infrastructure + craft. If your reply rate fell and you don't know why, this is the diagnostic. Audit honestly; the fix is almost never what the team thinks it is.

WHEN TO USE

Work through all 10 dimensions with whoever owns sending infrastructure (RevOps, SDR manager, or the unlucky person named in the last domain audit). Use the shared 1-5 legend. Don't assume infrastructure is fine because deliverability dashboards say 'green' — those are leading indicators at best. Total the score. Read the band recommendation. Fix infrastructure before craft; fix craft before volume.

Preamble. Cold email reply rates have collapsed from a 2019 blended average of ~8% to Instantly's 2026 benchmark of 3.43%. The teams still hitting 15-20% replies are not sending more email — they are sending better-targeted, better-configured, better-crafted email. This scorecard separates infrastructure leaks from craft leaks from volume mistakes. Most teams discover it is all three.

Source lineage:

- Instantly 2026 Cold Email Benchmark Report — 3.43% reply / 27.7% open baseline; 0.69-2.34% meeting rate (dimensions 1-5, 10)
- Google / Yahoo Feb 2024 sender guidelines — SPF + DKIM + DMARC now mandatory for bulk senders (dimension 1)
- Landbase 2026 intent-signal study — signal-specific personalization lifts reply ~47% and deal size ~43% (dimensions 6, 8)
- Schwartz Breakthrough Advertising — awareness-level matching at the body level (dimension 7)

Scoring legend (1-5). Every dimension uses the same 1-5 scale:

1. **1 — Absent:** You do not do this. Either you have not built it, you built it and retired it, or the discipline does not exist in the org.
2. **2 — Ad-hoc:** Someone does this, sometimes, when they remember. No cadence, no owner, no artifact. Cannot be audited.
3. **3 — Documented:** An owner exists. A cadence exists. An artifact exists. Not yet measured; not yet feeding other layers.
4. **4 — Measured:** Outputs are tracked against a baseline. Reviewed at least quarterly. Decisions are made from the numbers, not vibes.
5. **5 — Compounding:** Outputs feed back into upstream layers. Every cycle makes the next cycle sharper. The discipline survives the person who built it.

Total possible: 50. Your band determines where to look first.

Score range	Band	Where to look first
≤ 20	Foundation missing	Your infrastructure is compromised. Adding send volume or better copy now will make the problem worse. Stop sending for one week. Fix SPF/DKIM/DMARC, segregate onto cold-email-only domains, verify lists. Only then rebuild craft.
21 – 33	Leaking between layers	Infrastructure mostly OK, craft is the leak. Focus the next 30 days on signal-matched openers (dimension 6) and 3-signal minimum personalization (dimension 8). Hold volume flat while you fix the craft layer — sending more bad email accelerates reputation decay.
34 – 43	Working, not compounding	You are above the blended baseline but not yet at signal-qualified performance. The gap is usually dimensions 9 and 10 — testing discipline and benchmark vigilance. Install the A/B test log and monthly benchmark review; 12% → 18% reply rate is almost entirely a process-loop problem.

44 – 50	Compounding	You are in the top decile of B2B outbound. The risk now is complacency — domain reputation and warm-up maturity (dimensions 2-3) are the discipline that silently breaks at scale. Stress-test these quarterly; they take the longest to rebuild if they collapse.

DECISION CRITERIA

1. SPF / DKIM / DMARC configuration

The three authentication standards Gmail + Yahoo now enforce for bulk senders. A single misconfiguration drops you into Promotions at best, Spam at worst. DMARC p=quarantine with aspf=r is the minimum viable config for new programs; p=reject with alignment takes 2-4 weeks of monitoring to graduate into.

Score 1 (low): At least one of SPF / DKIM / DMARC is misconfigured, missing, or set to p=none with no monitoring.

Score 5 (high): All three pass on every sending domain. DMARC is at p=quarantine (min) with aggregate reports flowing to a monitored inbox. Aligned SPF + DKIM on the From domain.

Recommendation: Run mxtoolbox.com against every sending domain today. Set DMARC to p=quarantine with aspf=r within 48 hours. Monitor aggregate reports for 2 weeks before escalating to p=reject. Never jump to p=reject without the monitoring window — one misaligned service silently nukes a week of sends.

2. Domain reputation

The inbox providers score every sending domain independently. Google Postmaster Tools is the only truthful source for Gmail reputation; the "deliverability dashboards" in most sequencers are guesses. Reputation is binary in practice — once it collapses, the recovery window is measured in months.

Score 1 (low): Never checked Google Postmaster Tools. No domain reputation tracking. Sending from the primary marketing domain.

Score 5 (high): Postmaster Tools verified on every sending domain. Reputation reviewed weekly. Sending segregated onto cold-email-only domains (never the primary).

Recommendation: Verify every sending domain in Google Postmaster Tools this week. Segregate cold-email sends onto dedicated domains today — any overlap with your primary marketing domain is infrastructure risk you are taking on needlessly.

3. Warm-up maturity

Every cold-email domain needs 4-8 weeks of warm-up before touching real prospects. Most teams skip this, blast, and torch the domain in week two. Warm-up is not a tool you buy; it is a discipline of ramping volume against a responding audience.

Score 1 (low): No warm-up. Or "warm-up" = a Mailreef/Warmup Inbox auto-reply loop with no real human engagement.

Score 5 (high): 6+ weeks of progressive warm-up with a mix of internal-team replies and authentic engagement. Domain has sent <50/day for 30+ days before ramping to production volume.

Recommendation: If you are under 4 weeks of warm-up on any domain, cut the daily send volume on that domain to 30-50/day until week 6. The math on ramping too fast is brutal — one collapsed domain costs a month of deals.

4. Send volume ramp

Even a fully warmed domain should not jump from 50/day to 500/day overnight. Inbox providers pattern-match on velocity. A healthy ramp is +20-30% per week once the domain is producing replies at target rates.

Score 1 (low): Send volume jumps 3x+ week-over-week. Or volume is flat at the wrong level (too high for reply rate, too low for capacity).

Score 5 (high): Per-domain ramp schedule documented. Volume increases tie to actual reply-rate evidence, not "we need more meetings." Pause-and-reassess triggers at 30% reply-rate drop.

Recommendation: Write a per-domain ramp schedule (weeks 1-12) this week. Put a hard gate at each threshold: no ramp without a reply-rate floor. If reply rate drops 30% in a week, pause the ramp and diagnose before continuing.

5. List hygiene

Bounce rate above 3% starts degrading the domain within days. Catch-all and role-based addresses are the primary culprits. Spam-trap exposure is invisible until it is catastrophic — a single trap hit on a major blocklist can quarantine an entire IP block.

Score 1 (low): No verification step. Bounce rate unknown or >5%. Role-based addresses (info@, sales@) in the

Recommendation: Bounce rate above 3% is an emergency. Pull the list, re-verify, retire any address that has bounced once. If you don't know your bounce rate, you cannot claim the infrastructure is healthy.

6. Opener relevance (signal-matched vs generic)

The single highest-leverage craft variable. Landbase's 2026 data shows signal-specific personalization lifts reply rates ~47% over generic. "I see you are a VP of Sales" is not personalization — it is observed identity, which costs nothing and earns nothing.

Score 1 (low): Openers reference firmographics only ("I saw you are a VP at..."). No PSP signal referenced.

Score 5 (high): Every opener references a specific observed signal from the last 90 days — hiring pattern, public post, funding event, tech-install. The signal ties directly to the PSP definition.

Recommendation: Audit the last 100 openers shipped. Count how many reference a named recent signal vs. how many are identity-generic. If <50% are signal-matched, your craft layer is the leak — fix this before adding any new volume.

7. CTA clarity (one ask per email)

Multi-CTA emails convert worse than single-CTA emails. The math is not complicated: every additional option decreases the probability the prospect acts on any of them. Schwartz's awareness-level rule applies: match the CTA commitment to the awareness level of the prospect.

Score 1 (low): Emails have 2+ CTAs ("book a call OR reply OR grab our whitepaper"). CTA commitment level does not vary by awareness level.

Score 5 (high): Every email has exactly one CTA. CTA is calibrated to the awareness level — soft ask for unaware, specific ask for most-aware. Sequence choreographs escalation.

Recommendation: Rewrite every sequence so each email has one CTA. Audit the CTA commitment level across the sequence — if the first email asks for a 30-minute meeting, you are pricing the prospect out of reply. Start smaller and escalate.

8. Personalization depth (3+ signals)

One signal is noise; three signals is evidence. The threshold matters because it forces you to pause the send and ask whether the prospect is actually in a PSP — instead of shipping because the opener box is filled.

Score 1 (low): 0-1 signals per recipient. Personalization is {{first_name}} and company name.

Score 5 (high): 3+ signals referenced per recipient, each tied to the PSP. Sequence variants trigger on which signals are present (hiring vs content vs funding).

Recommendation: Set a hard rule: no send without 3 signals present. Yes, volume drops. Reply rate rises enough to compensate within one cycle. Teams that can hold this rule for a quarter stop having deliverability problems.

9. A/B test discipline

Most outbound "A/B tests" produce no decisions because they test multiple variables at once, run under-powered sample sizes, or get called based on a single day of data. The discipline is harder than the vocabulary.

Score 1 (low): A/B tests are vibes. Multiple variables per test. Sample sizes <500 per variant. Decisions made from first-day data.

Score 5 (high): One variable per test. Power calculations done in advance. Sample size met before calling. Learning logged in a shared doc. Losing variants retired immediately.

Recommendation: Write an A/B test log template this week. Every test requires a hypothesis, one variable, a sample-size target, and a decision date BEFORE launch. Tests that fail the template do not ship.

10. Reply-rate benchmark vs. Instantly 2026

Instantly's 2026 benchmark: 3.43% reply, 27.7% open, 0.69-2.34% meeting rate for the blended average. Teams running signal-qualified segmentation see 15-20% reply. If your reply rate is below 3.43% you are underperforming the blended baseline — which is usually an infrastructure or list-hygiene problem, not a craft problem.

Score 1 (low): Reply rate below 2%. Open rate below 20%. No benchmark comparison documented.

Score 5 (high): Reply rate ≥8% (above blended baseline). Open rate ≥35%. Benchmark comparison reviewed monthly. Underperformance triggers diagnostic, not more volume.