

EXAMPLE 03 · REPORTING

Six weeks. The threshold was missed.

Written as of 15 September 2026 · Scipioform

Synthetic demonstration — not client work or achieved results.

Stannary Systems is invented. It is not a Scipioform client, a disguised client, or a real company we know of; any resemblance to a real business is unintended. It exists so we can publish a complete piece of work instead of an excerpt, and so nothing here depends on a client giving us permission.

RESULT AGAINST THE WRITTEN THRESHOLD	
The pilot missed its threshold.	
The threshold was 6 qualified conversations, agreed in writing before sending began. It produced 4. That is a miss, and under the pilot terms the at-risk portion of the fee is refunded. Nothing about the rest of this report changes that, and the refund is not contingent on Stannary agreeing with our interpretation of why.	
THRESHOLD, AGREED IN WRITING BEFORE SENDING	PRODUCED
6 qualified conversations	4 (4 from 6 positive replies)

What "qualified" meant here

The word does not appear on any Scipioform surface without its definition. This is the definition agreed with Stannary Systems before any message was sent, with the dated clarification applied to every classification:

A conversation with a person who (a) holds or directly influences the budget for inspection software at an independent inspection contractor of 20–200 technicians, (b) confirms an active framework or term contract in mobilisation, and (c) agrees to a next step with a date on it. All three. Any one missing and it is not counted.

Clarification agreed 11 August 2026 (D-02): “in mobilisation” means the asset owner has issued a commencement date, whether or not work has begun. An award with no commencement date does not satisfy condition (b).

Why six. Six was set from what Stannary would need to see to justify continuing at the ongoing fee. It was not derived from a benchmark, and we did not have one to derive it from.

Window. Six weeks, 3 August to 11 September 2026.

The funnel, with every denominator

Each step says what it counts and, deliberately, what it does not. The second line is the one that usually gets left out, and it is the one that stops a number being read as more than it is.

204 Attempted sends 200 from 204 attempted sends	Messages the system tried to send. Not people: a few accounts received a follow-up in the same window.
200 Delivered 200 from 204 attempted sends	Accepted by the receiving server. Not "reached an inbox". We cannot see a spam folder, and we do not claim to. Three failed at rejected addresses; one domain no longer resolved.
18 Human replies 18 from 200 delivered	A reply written by a person. Not opens. We do not report opens: image-proxy prefetching makes them unreadable, and reporting a number we cannot interpret is worse than reporting none.
6 Positive replies 6 from 18 human replies	A reply that invites a next step. Not qualified. A positive reply is interest; it says nothing yet about budget or fit.
4 Met the written definition 4 from 6 positive replies	All three conditions above, evidenced in the thread or the call. Not opportunities. This is the number the threshold was set against.
3 Conversations actually held 3 from 4 met the written definition	A scheduled call that happened. The fourth was booked inside the window and sat outside it; it is counted as qualified and not as held. It is not counted twice.
1 Opportunity 1 from 3 conversations actually held	Stannary recorded an active deal with a next step and a date. Not revenue, not a forecast, and not a win. It is one deal at the earliest stage.

Reconciling the counts

Every number above has to account for the one above it. Where people go missing between two steps, this says where they went.

- 01 Attempted 204, delivered 200. Three messages failed at rejected addresses and one domain no longer resolved; and they stay in the attempted count rather than being quietly dropped.
- 02 Of 18 human replies, 6 invited a next step and 12 did not. Of those 12: 7 were declines, 3 were "not me, try this person" (all three were followed up and are counted once, at their first reply), and 2 are the ambiguity described below.
- 03 Of 6 positive replies, 4 met all three parts of the written definition. The 2 that did not: one was at a contractor of about 400 technicians, outside the segment; one could not confirm an active mobilisation.
- 04 Of 4 qualified, 3 conversations were held inside the window. The fourth was booked on 10 September for 16 September, which is outside it. It counts as qualified and not as held.
- 05 Of 3 held, 1 became an opportunity in Stannary's own records. The other two: one wants to revisit after its next award, one chose to extend its spreadsheet process.

Both variants, with their actual exposure

Synthetic allocation: 180 contacted accounts, 90 per variant. Eligibility was the intended filter; the failures below show that it was not enforced reliably. Each stayed in its assigned variant; 24 follow-up attempts bring the total to 204 messages. Four messages did not deliver, leaving 200 delivered. Reply counts are deduplicated by account. Assignment was balanced; the small outcome counts and possible differences between accounts still limit the comparison.

Invented teaching data: allocation, exposure and outcomes

Measure	Award named	Award not named
Accounts assigned	90	90
Messages attempted	102	102
Messages delivered	100	100
Human replies	12 from 100 delivered	6 from 100 delivered
Positive replies	5 from 100 delivered	1 from 100 delivered
Qualified conversations	3 from 100 delivered	1 from 100 delivered
Conversations held	2	1
Opportunities	1	0

On 26 August, a prospect exposed the unsupported evidence-format claim recorded as EX-02 in the rehearsal. Thirty-one accounts were suppressed from further sends; any messages already sent to them remain in the counts. Of 218 reviewed accounts, 180 were contacted and 38 were not. The suppression overlaps those groups: it is not another funnel step to subtract from delivered messages. The evidence

format for a fourth asset owner remained unverified at handover. Before another test, the client must resolve the product gap and confirm which claims are supportable. This mid-pilot list change also limits comparisons across weeks.

These are message-based response ratios, not probabilities for unique people. A held conversation is a subset of qualified conversations; an opportunity is a subset of held conversations. No outcome is added twice.

Where this report is uncertain

Two of the 18 replies read "send me some information". We classified both as not positive, because neither invited a next step and neither named a person or a date. That judgement is arguable. Had we classified them as positive, positive replies would read 8 from 18 rather than 6 from 18.

It would not change the number that matters. Neither of the two could have met the written definition without a further exchange that did not happen, so qualified stays at 4 and the pilot still misses. We are flagging it because the classification was a judgement we made, not a fact we observed, and because the same judgement applied at larger volume would matter.

The same six weeks, read two ways

This is the part worth taking away. Both readings below are arithmetically correct on the same six weeks, and they point in opposite directions.

THE READING MOST REPORTS LEAD WITH

Read on reply volume

18 replies from 200 delivered. Against a cold outbound programme into a defined segment, that is a healthy-looking response, and it is the number most reports would lead with.

Decision it leads to: Continue, and scale sending volume. The message is working; send more of it.

THE READING THE THRESHOLD WAS WRITTEN IN

Read on qualified conversations

4 from 6 positive replies met the definition, against a threshold of 6. The programme produced two-thirds of what it needed to.

Decision it leads to: Do not scale yet. Investigate the qualification failures before committing to more volume.

Both readings are arithmetically correct. They diverge because they measure different things: one counts attention, the other counts fit. Reply volume alone cannot establish that the programme should scale. Two of the six positive replies did not meet the written definition; the pilot missed the agreed threshold.

Scaling without investigating the qualification failures could consume more list budget and salesperson time without meeting the threshold. That is a risk the next test should examine, not a forecast proved by this pilot. That is the specific failure mode this firm reports against, and it is why the threshold is written in conversations that meet a definition rather than in replies.

The decision

DECISION RECORD

Do not scale sending. Resolve the evidence-format gap and verify eligibility, then propose four weeks with the same two message variants on the corrected list.

Of the 12 replies that did not invite a next step and the 2 positives that did not qualify, 5 came from companies outside the 20–200 technician band or with no active mobilisation. Those accounts should not have been in the list. That is an observed eligibility-check failure. It gives us a specific correction to test: corroborate the published award with a commencement date and a mobilisation signal. It does not establish that targeting was the only cause, or that the message could not improve.

With 100 delivered messages in each variant, the award-naming version produced 5 positive replies and 3 qualified conversations; the control produced 1 and 1. This is a small observed difference, not an established effect. Keep both variants in the proposed follow-up while correcting eligibility checks, then review qualified conversations against each variant’s actual delivered count.

OWNER

Nouman, Scipioform — list re-cut and the added verification step

CLIENT-SIDE OWNER

Stannary’s operations lead — confirm the technician-count band from the two accounts we got wrong

NEXT REVIEW

25 September 2026

FEE CONSEQUENCE

The at-risk portion is refunded for the missed threshold, on the terms agreed before the pilot started. The four weeks described above are a separate decision for Stannary to take or decline; they are not a way of working off the refund.

What we are deliberately not doing

- Keep the two messages stable for the proposed eligibility-check test. This limits simultaneous changes; it does not prove the copy is optimal.
- Do not increase volume before reviewing the corrected list and its resulting qualified conversations.
- Do not treat the trigger as established. The observed 3 versus 1 qualified conversations are too limited to justify that claim.

What this report does not claim

- No revenue is inferred. One early-stage opportunity is not a number.
- No causation is claimed. The observed difference supports a further test, without establishing what caused the outcome.
- No statistical significance is claimed or calculated. The counts and their denominators are shown so the reader can judge the limited evidence.
- These counts are invented for teaching. They are not Scipioform's results and not any client's.

This is the reporting standard we commit to on real engagements: every rate with its denominator, the misses reported alongside the rest, and a decision with a name and a date on it. How we will report results from now on (scipioform.com/work#reporting).