Startup Alignment Sprint

The evidence behind the method

Which findings carry my work, how solid they are, and where they stop.

Levi Neumann
M.Sc. Psychology · IHK-certified trainer · Aachen, Germany

Preface

Why this document exists

My website says I work on a psychological foundation. That is easy to claim and hard to verify. Hence this document.

It contains every study my approach rests on, with the full reference. You can look them up. You can hand them to someone who studied psychology. And you can hold me to them.

It also contains a section on what the research does **not** support. That one matters most to me. Anyone who promises you a percentage increase in revenue from team development has either not read the studies or is hoping that you won't.

> **How I rank evidence**
>
> **Meta-analysis**: combines many individual studies statistically and evens out chance findings. The strongest kind of evidence the social sciences have to offer.
>
> **Single study**: an indication, not proof. May come out differently in another context.
>
> **Internal company research or practitioner report**: makes something plausible without proving it. Wherever I lean on one, I say so.

Overview

The short version

What my offer claims, and how well each claim is supported. The last row is the most honest one.

Claim	Evidence	Strength
Teams where people can raise mistakes learn faster and perform better	Edmondson (1999); Frazier et al. (2017), meta-analysis of 136 samples	strong
Structured, facilitated debriefs improve performance	Tannenbaum & Cerasoli (2013), meta-analysis of 46 studies	strong
Decisive knowledge stays systematically unshared in groups	Lu et al. (2012), meta-analysis of 65 studies, 3,189 groups	strong
Contact between opposing groups works, but only under conditions	Pettigrew & Tropp (2006), meta-analysis of 515 studies	strong
If-then plans get carried out more often than statements of intent	Gollwitzer & Sheeran (2006), meta-analysis of 94 tests	strong
Contributions tied to a name prevent free riding	Karau & Williams (1993), meta-analysis of 78 studies	strong
Team development works, most strongly on collaboration and process	Klein et al. (2009), 60 correlations; Salas et al. (2008), 52 effect sizes, 1,563 teams	strong
Established teams benefit more than ad hoc ones	Salas et al. (2008), comparison of intact and ad hoc teams	moderate
Teams correct course at the midpoint of a project	Gersick (1988), qualitative study of eight teams	limited
Team development raises revenue or development speed	no solid basis, and I do not claim it	not supported

The sprint, phase by phase

What I do, and why

For each phase: what happens, which finding stands behind it, and where that finding stops.

Five phases, six weeks from kickoff to handover. The stated periods are the regular cadence. If holidays, a release or a funding round get in the way, we move them. The scope stays the same. That is not vagueness, it is the precondition for the sessions actually happening with the right people in the room.

Phase 1 Alignment Audit

approx. 1 week

What happens. A 90-minute deep dive with the founders and key roles. The result is a set of hypotheses, explicitly not a diagnosis.

Why they are only hypotheses. Leadership's picture is systematically incomplete at this point, and not because anyone is lying. In an interview study, 85 % of respondents reported situations in which they did not raise an issue they considered important with a superior (Milliken et al., 2003). Detert and Edmondson (2011) showed where that comes from: unwritten rules that nobody states and everybody follows. You don't contradict the boss in front of others. You need hard data before you open your mouth. You don't go over your direct manager's head.

What that means for the approach. For employees, silence is the rational choice. Ask only leadership and you get a clean, plausible, incomplete picture. That is why this phase ends with a list of hypotheses, which phase 2 tests against what the teams actually experience.

Limit: The 85 % come from 40 qualitative interviews. That is an order of magnitude, not a measurement, and that is how I use it.

Phase 2 Diagnostic Retrospectives

approx. 2 weeks

What happens. Two facilitated retrospectives, run separately by function, typically tech and customer-facing.

The single most important finding in this document. Tannenbaum and Cerasoli (2013) pooled 46 studies on structured debriefs. Teams that run them outperform comparable teams by roughly 25 % on average. The additional finding is what matters: **the effect is considerably larger when the debrief is structured, facilitated and aimed at specific improvements.**

That is exactly why I facilitate rather than hand the team a template. The difference between a facilitated retrospective and a feedback round is not a matter of taste. It is the difference between an effect and a wasted afternoon.

8×

less likely

That is how much less likely groups are to find the best solution when decisive knowledge sits with individuals. Not because people are bad at it, but because groups mostly discuss what everyone already knows.

Lu, Yuan & McLeod (2012): meta-analysis of 65 studies, 101 effects and 3,189 groups.

Why the protected setting is not a courtesy. For unshared knowledge to reach the table at all, you need psychological safety: the shared belief that you can admit a mistake or disagree without paying for it. Edmondson (1999) showed across 51 work teams that this factor acts on team performance through learning behaviour. The meta-analysis by Frazier et al. (2017), covering 136 samples and more than 22,000 people, confirmed the finding.

Limit: The 25 % is an average across very different fields, a lot of it medicine and the military. For a software team that is a reasoned expectation, not a prediction.

Phase 3 Management Insight Report

at the midpoint

What happens. The findings from the retrospectives are condensed into a report for the leadership team: about patterns and structures, not about people.

Why the report belongs in the middle. Gersick (1988) followed eight project teams across their entire lifespan, from seven days to six months. Teams do not reorient continuously. They do it in one jump, and almost exactly at the midpoint, **regardless of the absolute duration**. What matters is the relative position, not the date: the report comes at the midpoint, which in a six-week sprint means week three to four. If the schedule shifts, the report shifts with it. It does not arrive at some point, it arrives when course corrections tend to land.

Why patterns and not people. Kluger and DeNisi (1996) analysed 607 effect sizes on feedback. On average feedback helps, but **around 38 % of interventions made performance worse.** The difference lies in whether feedback targets the task or the person. A report that names names will very likely do damage.

Limit: Gersick is a qualitative study of eight teams. Well founded as a timing heuristic, not solid as a law.

Phase 4 Cross-Functional Action Sessions

approx. 1 week

What happens. Two sessions with deliberately mixed groups, ending in binding agreements.

Why mixed and not separate. Silos between tech and business are not a communication problem but an intergroup problem: two groups that typecast each other. Pettigrew and Tropp (2006) analysed 515 studies with 713 samples and 1,383 individual tests: contact between groups is consistently associated with lower prejudice, with a mean correlation of r = .21 in the expected direction. Contact only works under conditions, though: equal status in the situation, a shared superordinate goal, cooperation rather than competition. Those conditions do not arise on their own. They are the actual content of facilitation.

Why agreements are written as if-then plans. Gollwitzer and Sheeran (2006) pooled 94 tests with more than 8,000 people: intentions that specify **when**, **where** and **how** to act are carried out considerably more often than general intentions, with a medium to large effect (d = 0.65). “We want to communicate better” is therefore worthless. “If a ticket is blocked for more than two days, whoever owns it posts it in the channel” is not.

Why every commitment gets a name on it. Karau and Williams (1993) showed across 78 studies that individual effort drops in groups as soon as one's own contribution is no longer identifiable (g of about 0.44). A measure without an owner is a statement of intent.

Limit: Contact research comes mostly from societal intergroup contexts, not from companies. The mechanism transfers, the effect size does not one to one.

Phase 5 Performance Roadmap and Handover

approx. 1 week

What happens. A closing presentation to leadership and teams, the psychological roadmap for your scaling phase, the retro playbook with all templates, an assessment of who could facilitate your retrospectives from here on, and a follow-up check after three months.

Why the handover is the most important part. Leblanc et al. (2024) find a positive relationship between team reflexivity and team performance. It depends on boundary conditions, however, among them team size and tenure. So reflection does not work by itself. Learning research supplies the second half: Donovan and Radosevich (1999) analysed 63 studies and found spaced practice superior to massed practice, with the advantage shrinking as tasks get more complex. A one-off workshop is still an event, not a mechanism. That is why the sprint does not end with a presentation but with a procedure you can keep running yourselves.

Two independent meta-analyses reach the same conclusion. Klein et al. (2009) analyse 60 correlations, Salas et al. (2008) 52 effect sizes from 1,563 teams. Both find a moderate effect, and both find it **strongest for process outcomes**. In Salas et al., process is by far the best supported outcome type. Overall, team training explains 12 to 19 % of the differences between teams there; the authors note that this probably understates its practical significance.

An honest note: transfer into everyday work is the known weak spot of every people-development measure. The playbook, the host recommendation and the three-month check cushion it but do not solve it. Anyone who wants to run retrospectives permanently in-house needs people who can facilitate, and that is not learned from a document. There is a separate format for it; it is deliberately not part of the sprint, so that nobody here pays for something they may not need.

Scope

Which phase belongs to which offer

You don't have to start with everything. The building blocks are the same, they are only cut to different lengths.

	Alignment Diagnosis	Startup Alignment Sprint	Retro Host Training
Phase 1 – Alignment Audit	•	•	–
Phase 2 – Diagnostic Retrospectives	•	•	–
Phase 3 – Management Insight Report	•	•	–
Phase 4 – Cross-Functional Action Sessions	–	•	–
Phase 5 – Roadmap and Handover	–	•	–
Training internal facilitators	–	–	•
Impact measurement before and after	•	•	–
Follow-up check after three months	–	•	–
Duration	2–3 weeks	6 weeks	2 days

The diagnosis is the beginning of the sprint. Anyone who scales up later pays no more in total. The training is not part of the sprint but what comes after it: the sprint gives you the procedure, the training gives you people who can run it.

Transferability

Does any of this apply to your team?

The fair question about any study is: fine, but was this studied on teams like ours? The honest answer has a good half and an uncomfortable one.

The large meta-analysis by Salas et al. (2008) allows a more precise answer here, because it asks not only **whether** team training works but **where it works better and where worse**.

Three results speak for your case

- **Established teams benefit more than ad hoc ones.** For performance outcomes, the effect for intact teams sits clearly above the one for ad hoc groups, with confidence intervals that do not overlap. That is notable, because a large share of team research works with short-lived groups. Yours is the case where it works better.
- **Training that targets collaboration beats training that targets the task**, at least for process and climate outcomes. That is the difference between my work and another round of process optimisation.
- **More content is not better.** Programmes mixing task and teamwork content were not superior to focused ones. The authors suspect overload. That is why the sprint does few things rather than many.

And the uncomfortable half

The authors write themselves that the large majority of the studies analysed were run with students in simulated settings, a consequence of decades of funding by military institutions. Military settings reach the highest values, lab studies the middle, medical ones the lowest.

> **The number I do not use**
>
> The same meta-analysis reports a relationship of **.85** for studies from **business contexts**, against **.34** for the full dataset. That is by far the highest value in the whole analysis, and it rests on five studies. The authors explicitly warn against interpreting effect sizes based on so few studies.
>
> It would be the most flattering number I could quote, and it comes from exactly the context you work in. I am naming it here rather than in a sales conversation, because a dossier that shows only the convenient findings would not be one. I will not use it.

What follows from that: the mechanisms are well supported and transferable, meaning structured debriefs, psychological safety, shared knowledge, binding intentions. The exact numbers are not. For a 40-person software company in Aachen there is no study. There are well-founded expectations, and there is a willingness to check after three months whether they came true.

A finding against the grain

Why I don't let people “just talk it out”

A widespread piece of management wisdom says: conflict about the issue is good, conflict about people is bad. The first is even said to be productive. That distinction appears in many textbooks.

It does not hold up. De Dreu and Weingart (2003) pooled the available studies and find that **both** types of conflict correlate negatively with performance and satisfaction, more strongly so for complex tasks. The authors explicitly note that their result contradicts what both the research literature and introductory textbooks suggest.

> **What that means in practice**
>
> “Let's just talk it out openly” is not a method, it is the absence of one. Task conflict does not become productive because you allow it, but because the order, the right to speak and the output format are fixed beforehand. That is why I work with a firm structure rather than an open round.

And the limit of my method

Friedrich Glasl described conflict escalation in nine stages. From roughly the fourth stage on, the parties are no longer interested in the issue but in the other side losing. Past that point, facilitation is structurally out of its depth. Then it takes mediation, and in some cases a personnel decision.

I check for this in the first conversation. If I conclude that you need something other than what I offer, I say so before any money changes hands.

The counter-check

What the research does not support

An evidence dossier is worth only as much as the section in which its author argues against himself. Here it is.

- **No causal claim about business figures.** None of the meta-analyses cited shows that team development raises revenue, margin or development speed. What is supported are effects on collaboration, process and climate. That these act indirectly on business figures is plausible, but plausibility is not proof. What is supported is precisely the part you are hiring me for. The second half of the chain you can check yourselves, because you have your numbers and I don't.
- **No percentage promises for your individual case.** The 25 % from the debrief meta-analysis is an average across 46 studies. An average says nothing about where your team will land. That is why I measure with you instead of promising.
- **Meta-analyses average over contexts that are rarely startups.** This is not my objection but the authors' own observation: most of the studies analysed work with students in simulated settings, shaped by military research funding. The mechanisms transfer, the numbers only in part.

— **Small numbers of studies produce large numbers.** The narrower you cut a subset, the more spectacular the effects look and the less they mean. Effect sizes from fewer than five studies are not interpretable, not even when I like them. The findings I rely on here rest on dozens to more than a hundred individual studies.

— **Google's “Project Aristotle” is not a study.** The internal analysis of 180 teams that identified psychological safety as the most important factor is neither peer reviewed nor published. I mention it because it is well known and points in the same direction. What carries the argument is Edmondson and Frazier et al., not Google.

— **Publication bias.** Studies that find an effect get published more often than studies that don't. The reported effects are therefore closer to an upper than a lower bound. That applies to the entire literature, including the work cited here. What it shifts is the size of the effects, not their direction. That is why I argue with direction rather than with decimal places.

— **I do not promise a result.** I promise an approach that matches what research has shown to work, and an honest read of the situation when something does not.

Your own evidence

Whether it worked for you, I measure

The previous page says there is no study for a 40-person software company in Aachen. That is true. Which is why I generate numbers for your case.

What is measured. Psychological safety, the single best supported characteristic of teams that work and precisely the factor the sprint operates on. It is no substitute for a result, but it is the most solid available indicator of whether something has moved in how you work together.

With what. The **PsySafety-Check** by Fischer and Hüttermann, a short instrument validated for German-speaking use and published in the GESIS Leibniz Institute's collection of social science items and scales. Seven questions, three minutes, anonymous. For English-speaking teams, the original Edmondson scale from which the German questionnaire is derived.

When. Once before the first retrospective, a second time at the follow-up check after three months. **Deliberately not right after the sprint:** a value taken immediately after a workshop is regularly inflated: fresh impressions, good mood, the wish that it was worth something. The interesting value is the one that survives three months of everyday work.

What I will say about it, and what I won't

I will say: “The value rose from 4.1 to 5.3.” That is a describable change, and you can follow how it was arrived at.

I will **not** say “significantly improved”. With eight to fifteen people there is no sound basis for inferential statistics. Anyone arguing significance at that group size is hoping nobody asks.

That is not a weakness of the method but a question of the right tool. Inferential statistics answers whether a finding holds beyond the group studied, whether it would apply to other teams as well. That question does not arise here: this is about your team, not about teams in general. For that, description is not a weaker substitute but the right instrument.

And if nothing has moved, I say that too. Then we talk about why.

What you get out of it, independent of me. You get a baseline for your team that you can measure again later, using a freely available, documented instrument. That stays with you, even if we never work together again.

The survey is anonymous; I do not report results separately for groups of fewer than five people. Whether I may use the values in anonymised form, with no way back to your company, for my own analyses is something we put in writing beforehand, or we don't, and then they stay with you.

References

Source list

Every entry is written so that you can look it up directly.

De Dreu, C. K. W., & Weingart, L. R. (2003). Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. **Journal of Applied Psychology, 88**(4), 741–749.

Detert, J. R., & Edmondson, A. C. (2011). Implicit voice theories: Taken-for-granted rules of self-censorship at work. **Academy of Management Journal, 54**(3), 461–488.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. **Journal of Applied Psychology, 84**(5), 795–805.

Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. **Administrative Science Quarterly, 44**(2), 350–383.

Fischer, N., & Hüttermann, H. (2020). PsySafety-Check (PS-C): Fragebogen zur Messung psychologischer Sicherheit in Teams. In **Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)**. GESIS – Leibniz Institute for the Social Sciences.

Frazier, M. L., Fainshmidt, S., Klinger, R. L., Pezeshkan, A., & Vracheva, V. (2017). Psychological safety: A meta-analytic review and extension. **Personnel Psychology, 70**(1), 113–165.

Gersick, C. J. G. (1988). Time and transition in work teams: Toward a new model of group development. **Academy of Management Journal, 31**(1), 9–41.

Glasl, F. (2013). **Konfliktmanagement. Ein Handbuch für Führung, Beratung und Mediation** (11th ed.). Bern: Haupt. (First edition 1980)

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. **Advances in Experimental Social Psychology, 38**, 69–119.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. **Journal of Personality and Social Psychology, 65**(4), 681–706.

Klein, C., DiazGranados, D., Salas, E., Le, H., Burke, C. S., Lyons, R., & Goodwin, G. F. (2009). Does team building work? **Small Group Research, 40**(2), 181–222.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. **Psychological Bulletin, 119**(2), 254–284.

Leblanc, P.-M., Harvey, J.-F., & Rousseau, V. (2024). A meta-analysis of team reflexivity: Antecedents, outcomes, and boundary conditions. **Human Resource Management Review, 34**(4), 101042.

Lu, L., Yuan, Y. C., & McLeod, P. L. (2012). Twenty-five years of hidden profiles in group decision making: A meta-analysis. **Personality and Social Psychology Review, 16**(1), 54–75.

Milliken, F. J., Morrison, E. W., & Hewlin, P. F. (2003). An exploratory study of employee silence: Issues that employees don't communicate upward and why. **Journal of Management Studies, 40**(6), 1453–1476.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. **Journal of Personality and Social Psychology, 90**(5), 751–783.

Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. **Human Factors, 50**(6), 903–933.

Tannenbaum, S. I., & Cerasoli, C. P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. **Human Factors, 55**(1), 231–245.

A note on my own account

One frequently quoted number is deliberately missing from this document: that 23 % of failed startups name “the wrong team” as the cause. It comes from founders' self-reports in hindsight, not from a controlled study, and people rarely explain their own failure with themselves. Usable as a hook, not as evidence.

Questions about a particular finding?

Then write to me. I would rather argue about a study than about a brochure.

Levi Neumann · M.Sc. Psychology · IHK-certified trainer

anfrage@levineumann.de · levineumann.de · Aachen, Germany