ISD 2026 · Prague · Track T3 · 2 September 2026

From Tool Adoption to Organisational Design

An Empirical Study of AI Integration in Agile Software Development

Andreas Hinderks · Jörg Thomaschewski · Eva-Maria Schön

Hannover University of Applied Sciences and Arts · University of Applied Sciences Emden/Leer

jan-andreas.hinderks@hs-hannover.de · hinderks.org

Section 1

Tool or organisation?

Is AI integration something a developer adopts, or something an organisation has to design?

AI coding assistants are used daily
by most professional developers.
Organisations still treat them as individual
tools within stable team structures.

Why the tool-centric view is insufficient

- **73.3 %** of practitioners identify fragmented tooling as their primary frustration.
- Individual satisfaction with AI assistants rises **while team-level collaboration and knowledge sharing decline.**
- Sociotechnical systems theory predicts exactly these symptoms: technology integration requires **attention to institutions, culture and governance.**

Over 87 % of developers use AI tools daily; controlled experiments report productivity gains of 20 to 50 % on bounded tasks.

Herda et al. (2025) · Wivestad et al. (2024) · Kudina & van de Poel (2024) · Peng et al. (2023) · Ziegler et al. (2024)

The proposal under test

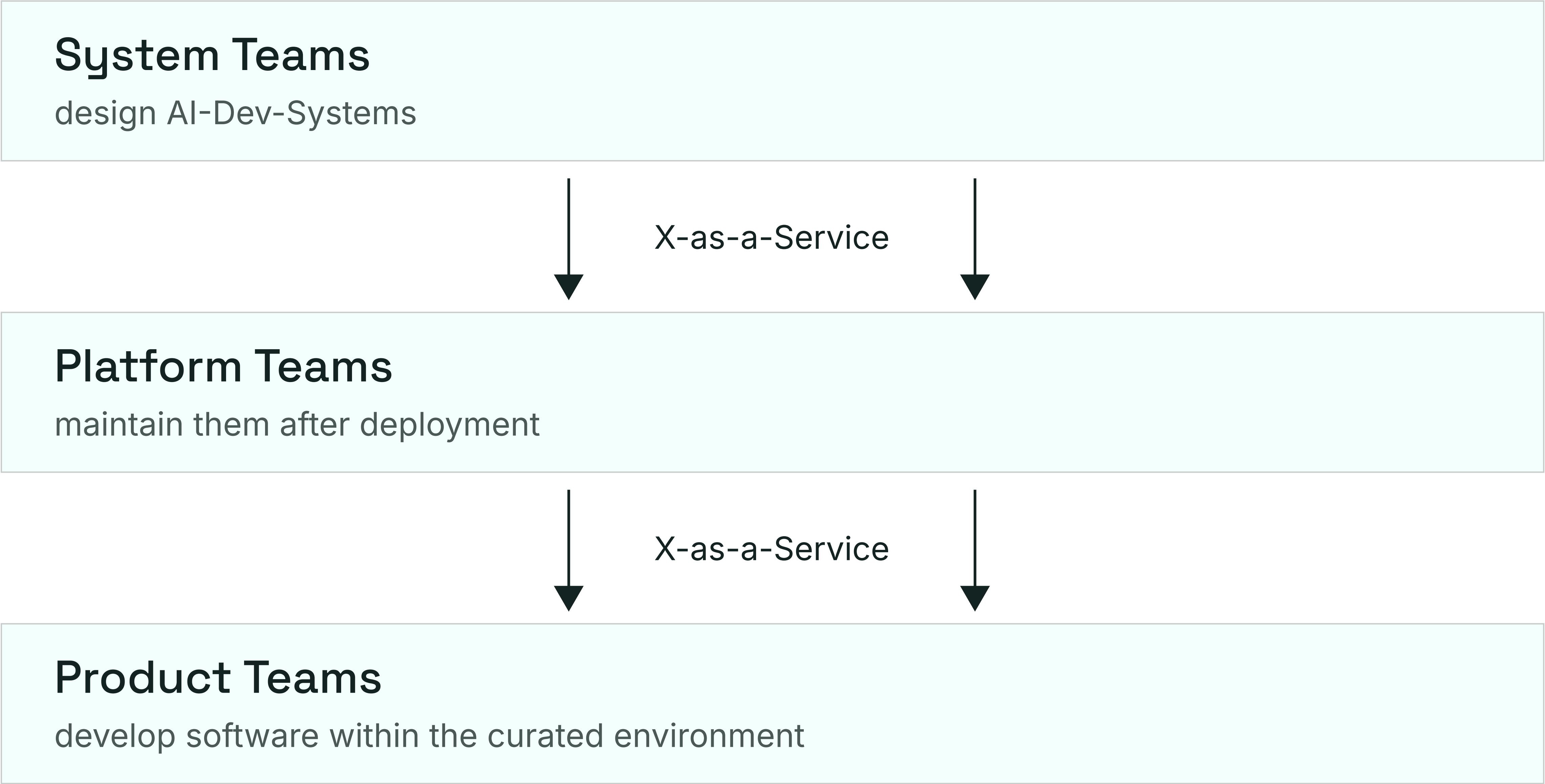


Fig. 1: The three-layer model: System Teams design, Platform Teams maintain, and Product Teams operate within the AI-Dev-System (Hinderks et al., 2026). Redrawn from the original figure.

An *AI-Dev-System* is a configured environment of AI agents, workflows and knowledge bases, treated as shared infrastructure.

The model extends the platform-team pattern of Team Topologies, whose central idea is to minimise **cognitive load** by abstracting complexity into platforms consumed through well-defined interfaces.

Proposed in our XP 2026 position paper. Conceptual, no data. This study supplies the test.

Research questions

- **RQ1** · To what extent do practitioners perceive AI integration as an organisational design problem rather than a tool adoption problem?
- **RQ2** · How do they assess the three-layer model's plausibility?
- **RQ3** · What changes in roles, iterations, and agile practices are they experiencing?

Each research question is operationalised by one measurement block:

RQ1 → Block O

RQ2 → Block M

RQ3 → Blocks C and A

The four hypotheses on the next slide follow from the same source.

Four hypotheses

	Hypothesis	Tested on
H1 RQ1	Practitioners who use AI tools more frequently rate AI integration significantly more often as an organisational problem than a tool adoption problem.	Block O mean by AI usage frequency
H2 RQ2	Practitioners working in large teams (≥ 10 persons) rate the three-layer model as more plausible than practitioners in small teams.	Block M mean by team size
H3 RQ1	There is a positive correlation between perceived fragmentation of AI tools in the team and agreement with the need for organisational restructuring.	C6 against the structural-need composite (O1, O3, O5)
H4 RQ3	Practitioners who use AI tools intensively report significantly more often a shift in their role from “producer” to “steward”.	C3 and C5 by AI usage frequency

Each hypothesis follows from one of the four propositions of the position paper — the same source as the four measurement blocks.

H1, H2 and H4 are group comparisons. H3 is a correlation on the full sample.

Section 2

Method

Instrument

- **28 Likert items** in four blocks: C perceived changes, O organisational challenge, M model plausibility, A agile practices.
- **Blocks defined a priori** from the four propositions of the position paper, not extracted afterwards by factor analysis.
- **7-point scale** normalised to −3 … +3, plus one open-text question.
- **Interpretation thresholds:** Strong ≥ +1.5 · Moderate +0.5 to +1.5 · Neutral −0.5 to +0.5 · Disagreement < −0.5.

A priori grouping ensures content validity by design, but renders the blocks thematic item groups rather than reflective scales. Interpretation therefore proceeds at item level, supported by item-total correlations; coherent sub-clusters are C1–C5, O1–O3 and M1–M5.

Welch's t where both groups satisfy Shapiro-Wilk, Mann-Whitney U otherwise, Bonferroni-corrected. Spearman's ρ for correlations. Effect sizes: Cohen's d and rank-biserial r.

Item wordings are paraphrased throughout this talk.

Sample

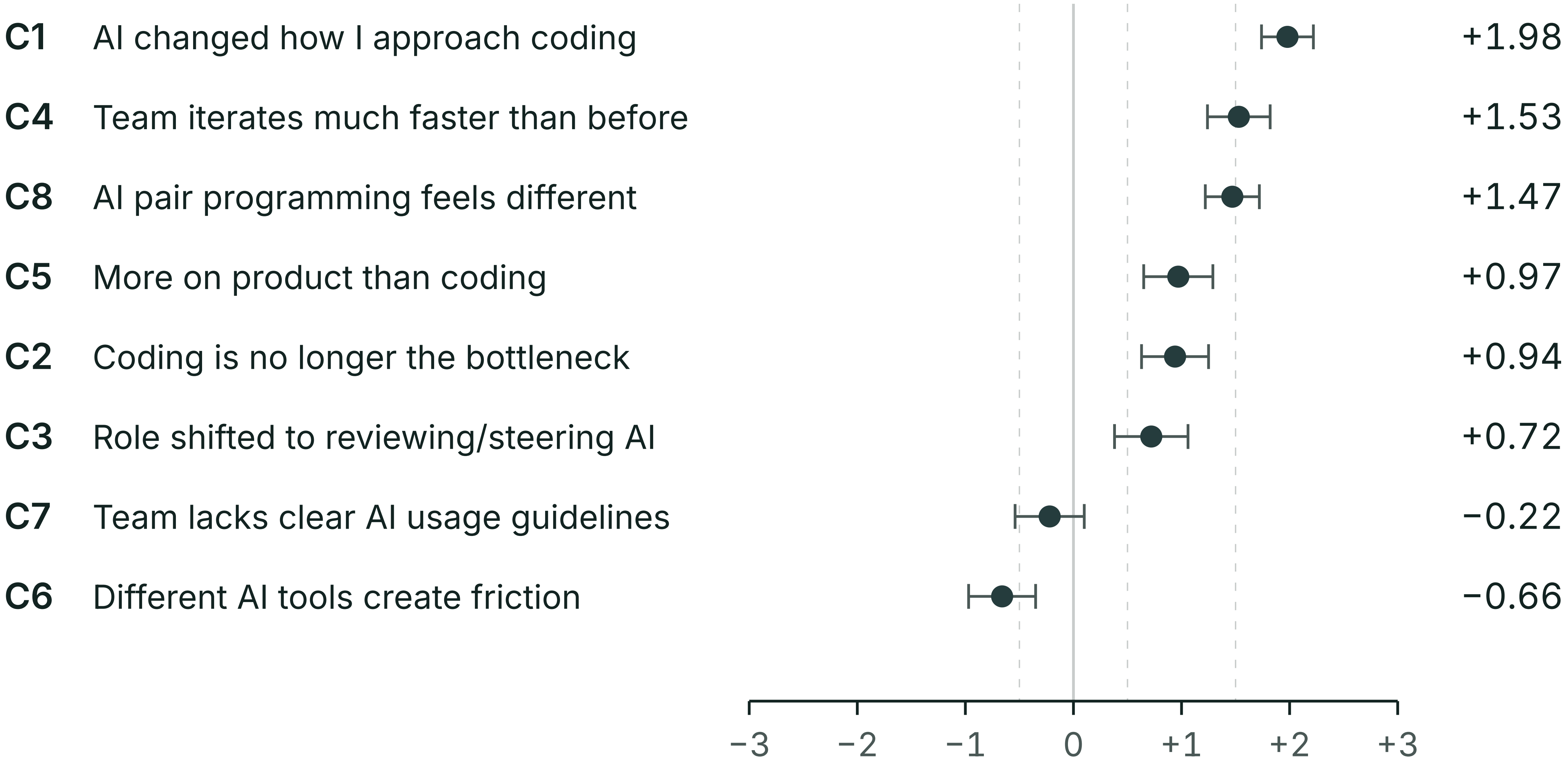
- **n = 99** professional developers, recruited via Prolific in **March 2026.**
- Three screening criteria: professional developer, experience with agile methods, at least occasional use of AI-based coding tools.
- **87 (88 %)** report daily or frequent AI tool use, **12** occasional or rare.
- **88 (89 %)** work in teams of 1 to 10 persons, **11** in teams of 11 or more.

A screened convenience sample; both grouping variables are imbalanced, which constrains H1, H2 and H4.

Section 3

Findings

Change is already under way — fragmentation is not yet felt



Strong ≥ +1.5 · Moderate +0.5 to +1.5 · Neutral −0.5 to +0.5 · Disagreement < −0.5

Fig. 2: Block C, perceived changes from AI tools. Item means with 95 % confidence intervals; this study, n = 99, scale −3 … +3.

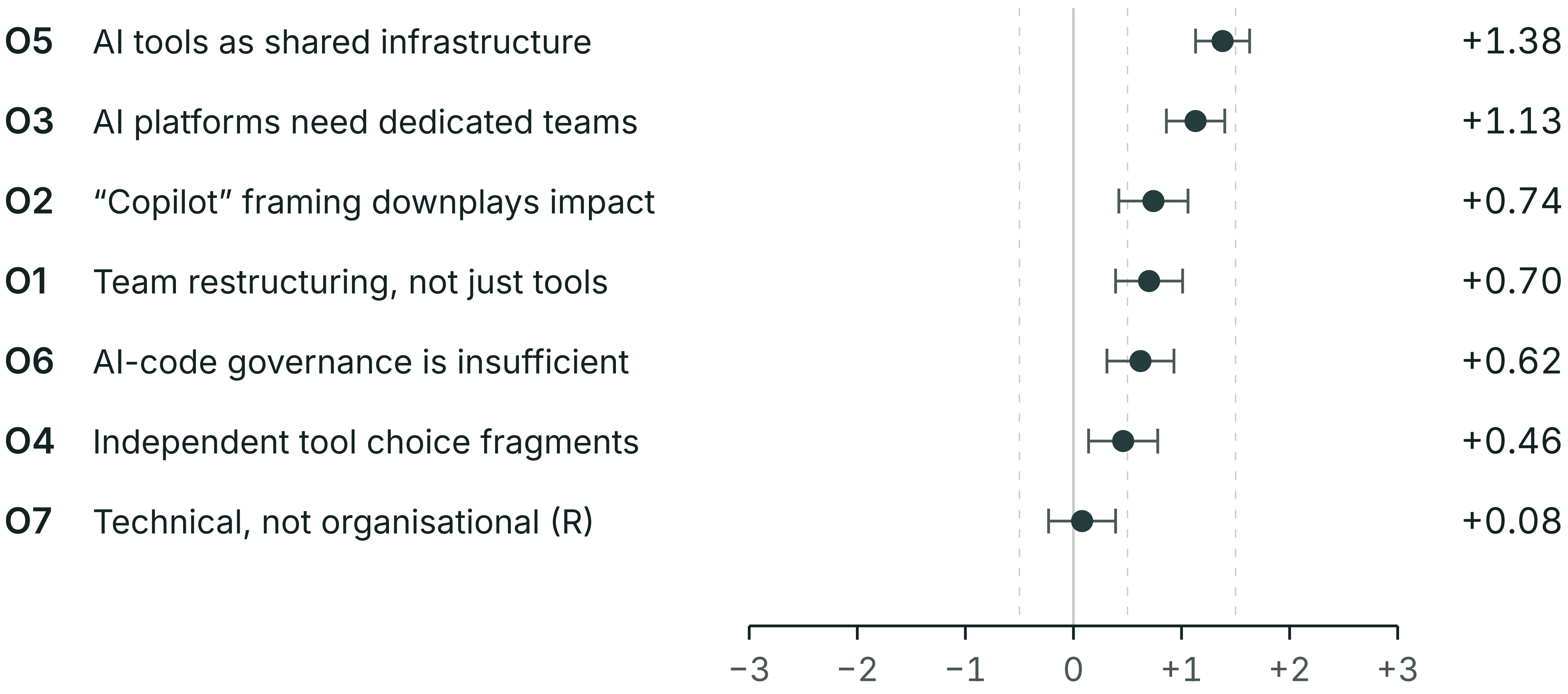
Block C · RQ3 · sets up H4, and C6 feeds H3

C1 and C4 record near-consensus: AI has changed how coding is approached, and teams iterate faster.

C3 and C5 give the **first evidence of the producer-to-steward shift** — H4 tests it.

C6 and C7 sit at the bottom: fragmentation and missing guidelines are not yet widely felt. That makes C6 useful as an early indicator, not as a complaint.

Practitioners endorse the infrastructure principle



Strong ≥ +1.5 · Moderate +0.5 to +1.5 · Neutral −0.5 to +0.5 · Disagreement < −0.5

Fig. 3: Block O, AI integration as an organisational challenge. Item means with 95 % confidence intervals; this study, n = 99, scale −3 ... +3. O7 is reverse-coded.

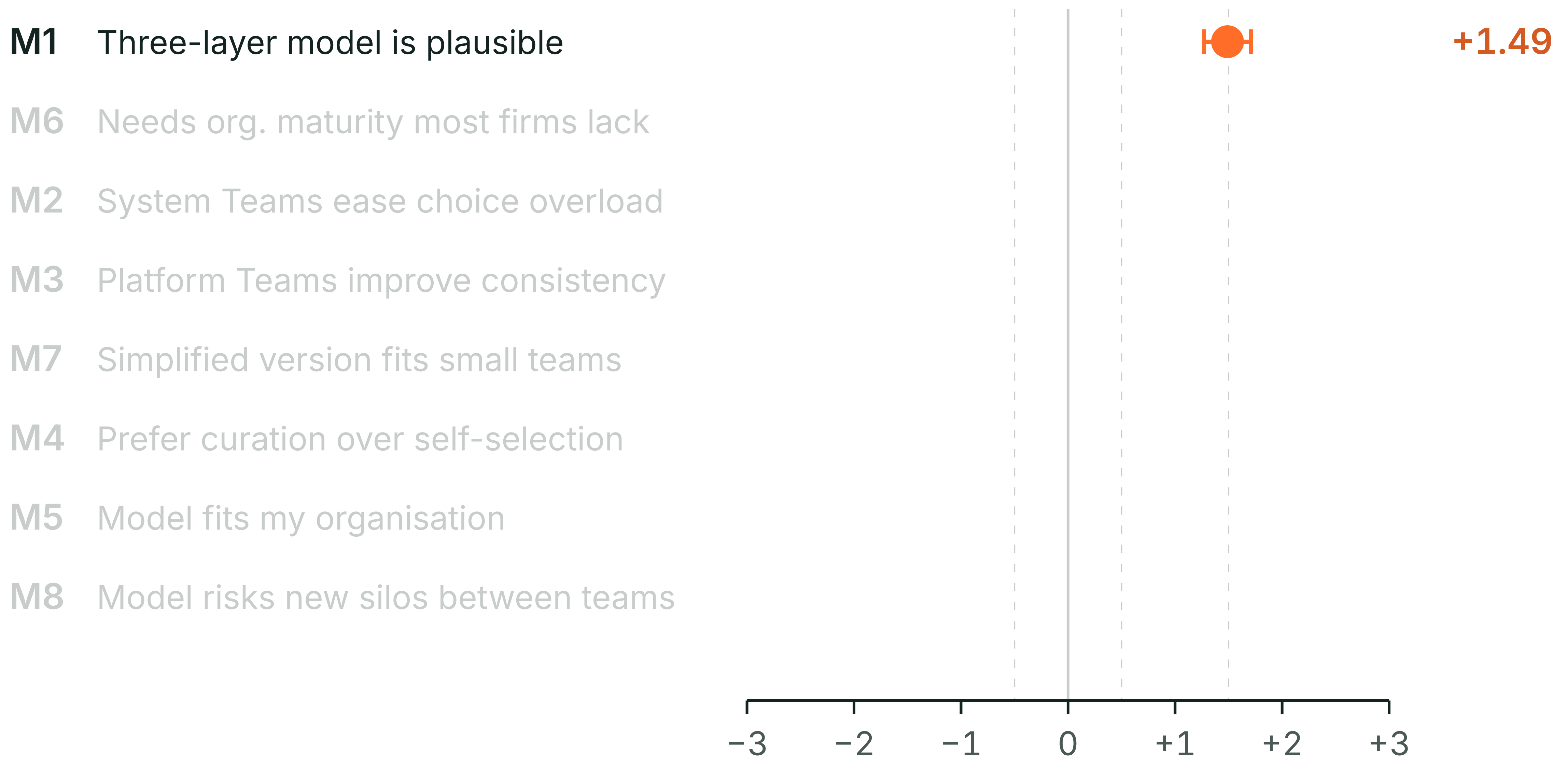
Block O · RQ1 · tested by H1 and H3

The two strongest signals are **O5 shared infrastructure** and **O3 dedicated teams** — qualified support for the organisational reframing.

Yet O4 is Neutral: fragmentation from independent tool choice is not acutely felt. And the confidence interval of the reverse-coded O7 straddles zero — no consensus on whether the challenge is technical or organisational.

Practitioners endorse the infrastructure principle before experiencing the pain that would motivate it.

A plausibility-to-applicability gap

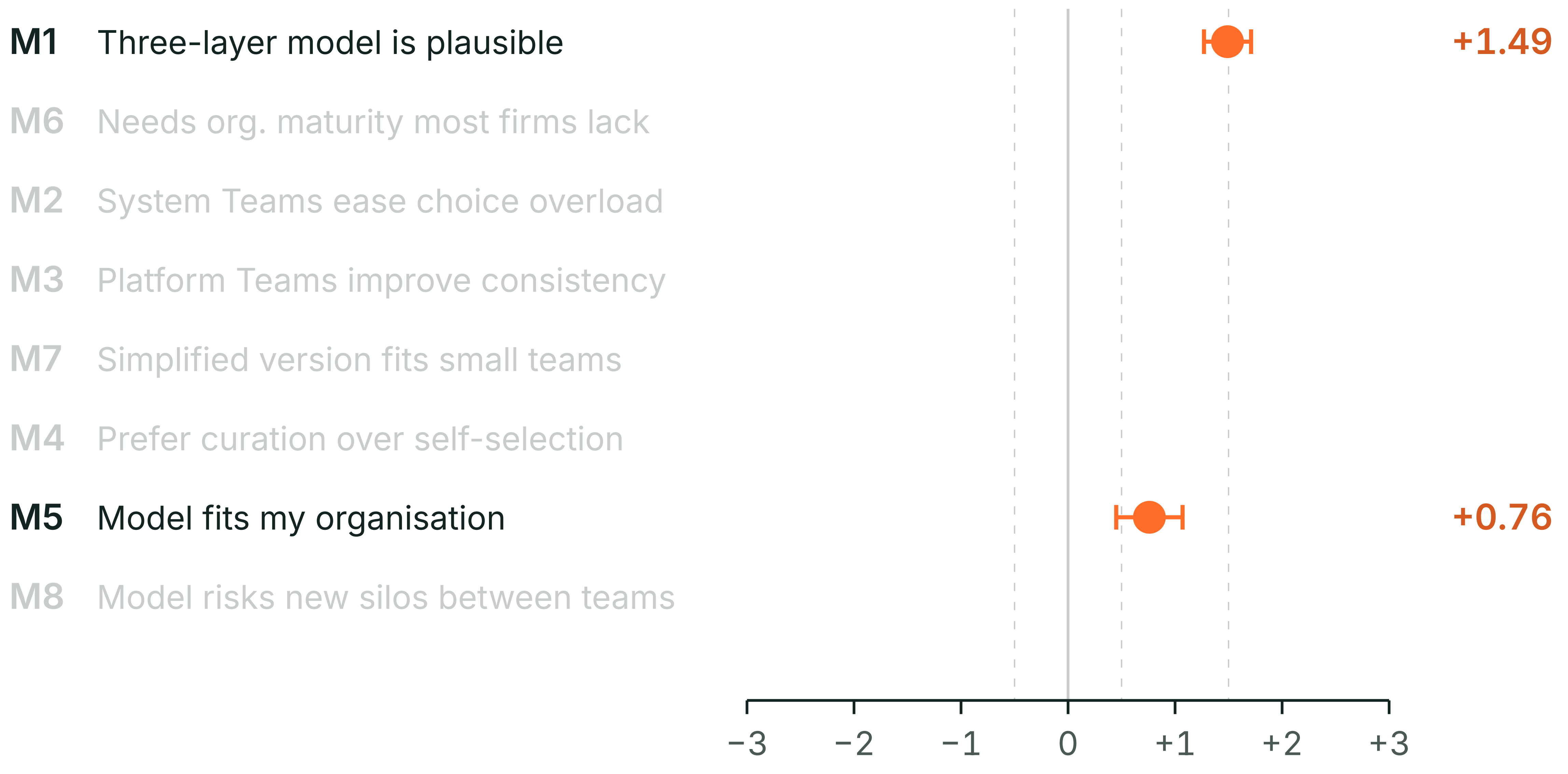


Block M · RQ2 · tested by H2

Strong ≥ +1.5 · Moderate +0.5 to +1.5 · Neutral −0.5 to +0.5 · Disagreement < −0.5

Fig. 4: Block M, three-layer model plausibility. Item means with 95 % confidence intervals; this study, n = 99, scale −3 ... +3.

A plausibility-to-applicability gap



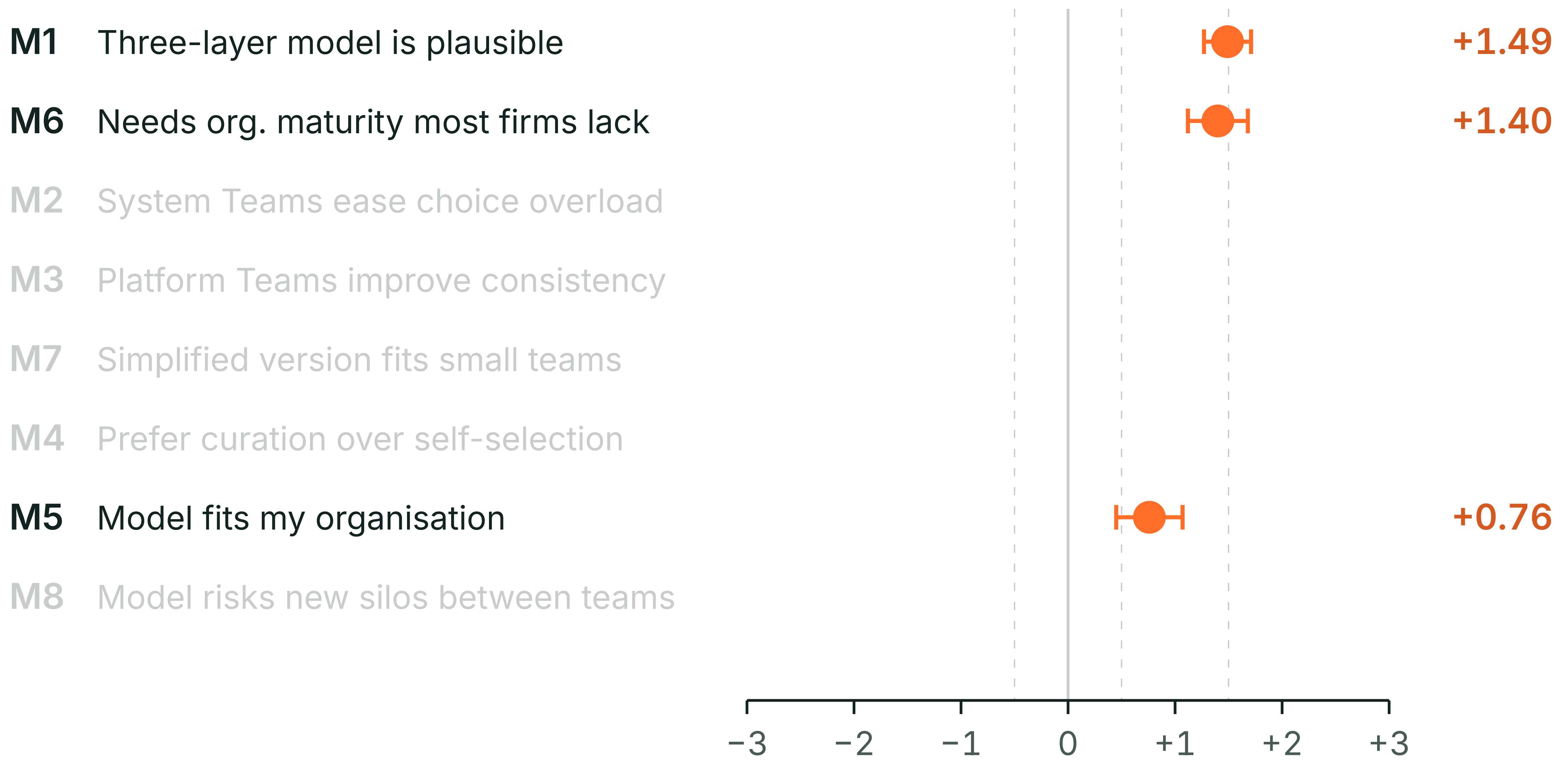
Strong ≥ +1.5 · Moderate +0.5 to +1.5 · Neutral −0.5 to +0.5 · Disagreement < −0.5

Block M · RQ2 · tested by H2

M1 → M5. Abstract plausibility is high. It drops markedly once respondents consider their own organisation.

Fig. 4: Block M, three-layer model plausibility. Item means with 95 % confidence intervals; this study, n = 99, scale −3 ... +3.

A plausibility-to-applicability gap



Strong ≥ +1.5 · Moderate +0.5 to +1.5 · Neutral −0.5 to +0.5 · Disagreement < −0.5

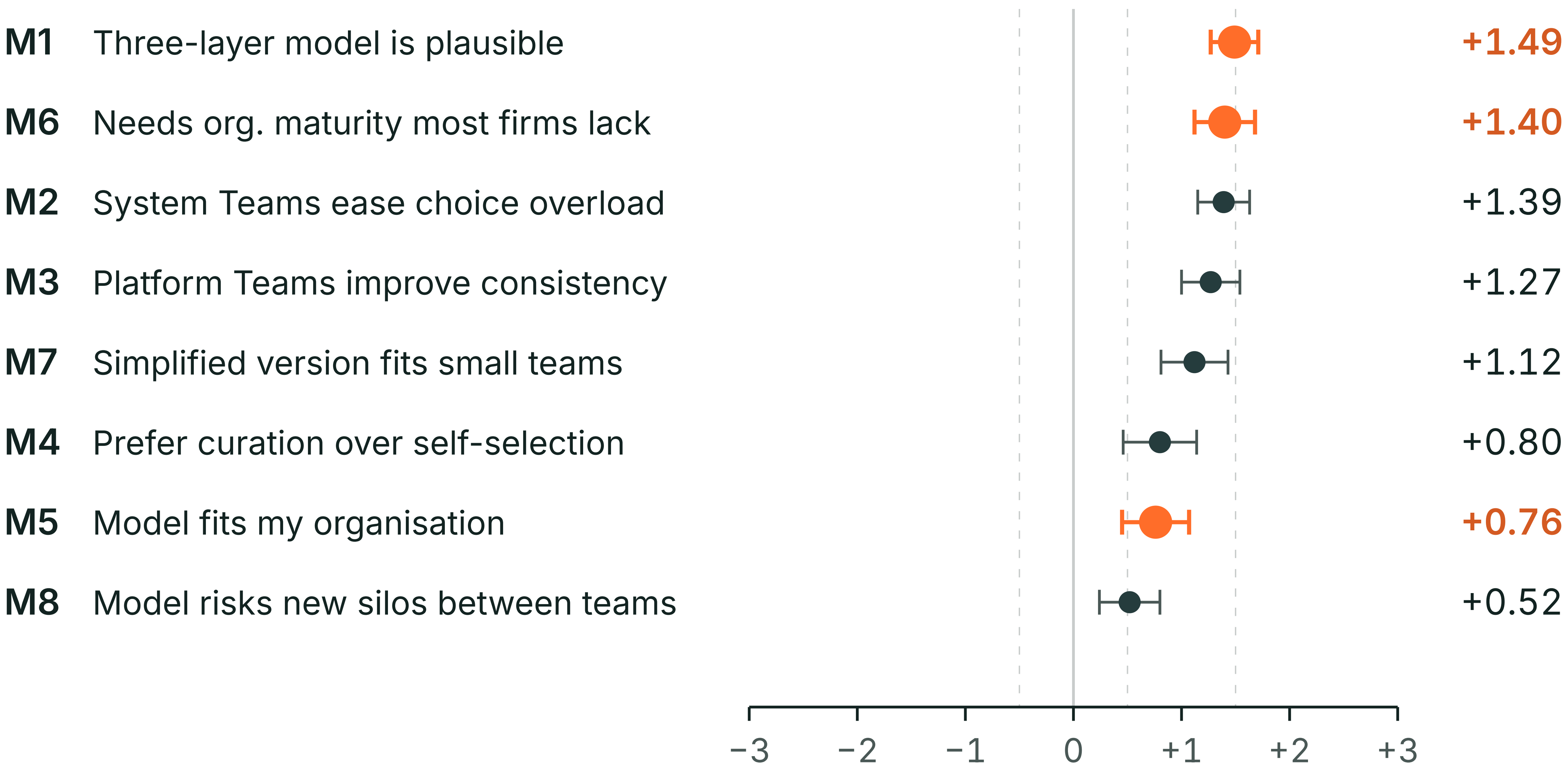
Block M · RQ2 · tested by H2

M1 → M5. Abstract plausibility is high. It drops markedly once respondents consider their own organisation.

M6 supplies the cause: the model demands organisational maturity most companies lack.

Fig. 4: Block M, three-layer model plausibility. Item means with 95 % confidence intervals; this study, n = 99, scale −3 ... +3.

A plausibility-to-applicability gap



Strong ≥ +1.5 · Moderate +0.5 to +1.5 · Neutral −0.5 to +0.5 · Disagreement < −0.5

Fig. 4: Block M, three-layer model plausibility. Item means with 95 % confidence intervals; this study, n = 99, scale −3 ... +3.

Block M · RQ2 · tested by H2

M1 → M5. Abstract plausibility is high. It drops markedly once respondents consider their own organisation.

M6 supplies the cause: the model demands organisational maturity most companies lack.

M7 offers a path — a simplified version for small teams. M8 registers borderline concern about new silos.

Hypothesis tests

H1 Block O mean by AI usage frequency

Welch's t, $p < .001$, $d = 1.63$ — supported

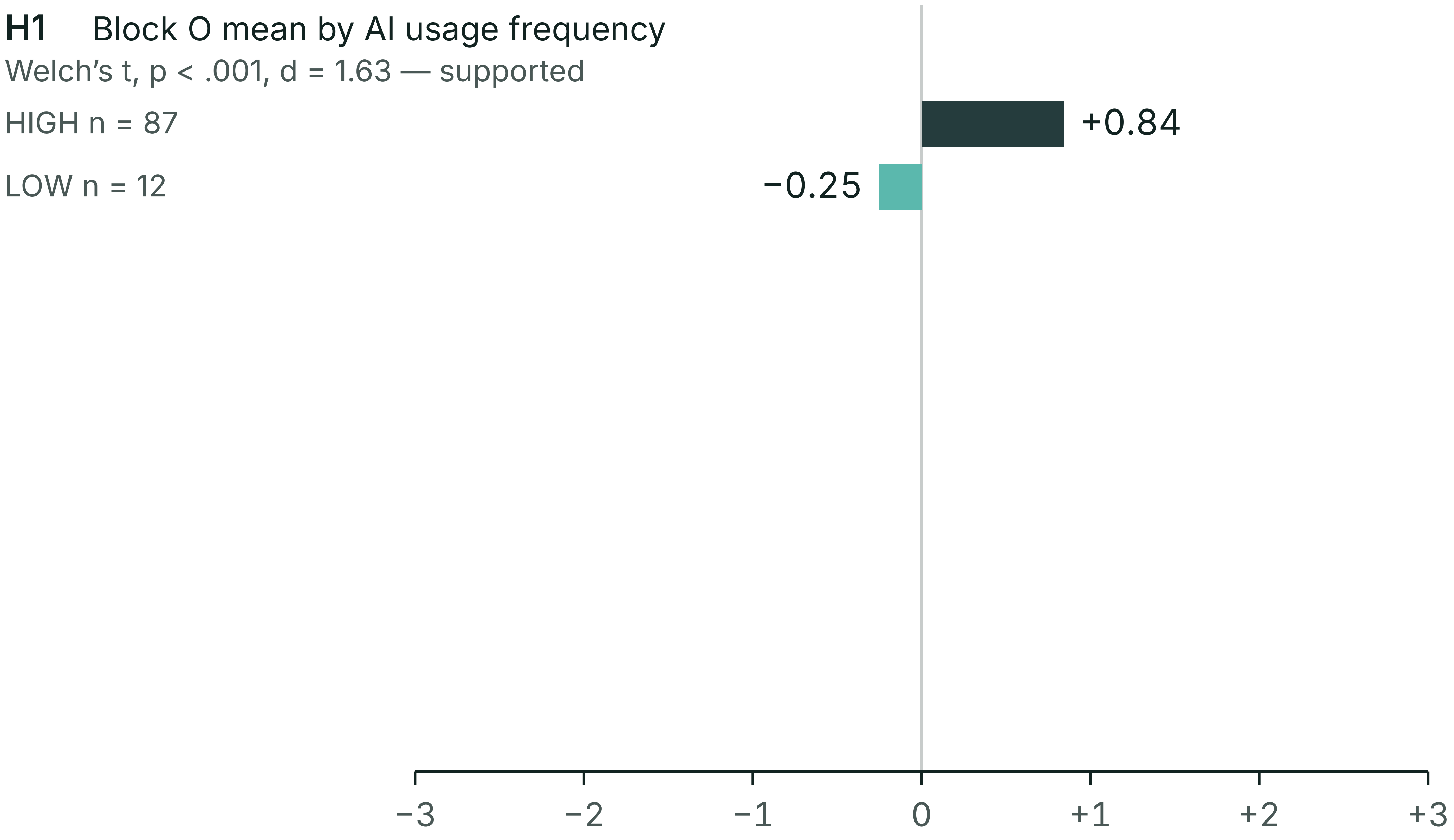


Fig. 5: H1, H4a and H4b by AI usage frequency. HIGH = daily or frequent use, n = 87; LOW = occasional or rare use, n = 12. Group means without CI — the paper reports none for the groups. This study, n = 99, scale −3 … +3.

H1 to H4 · the four hypotheses from Section 1

H1 supported. The more intensively practitioners use AI, the less it looks like a tool problem.

Hypothesis tests

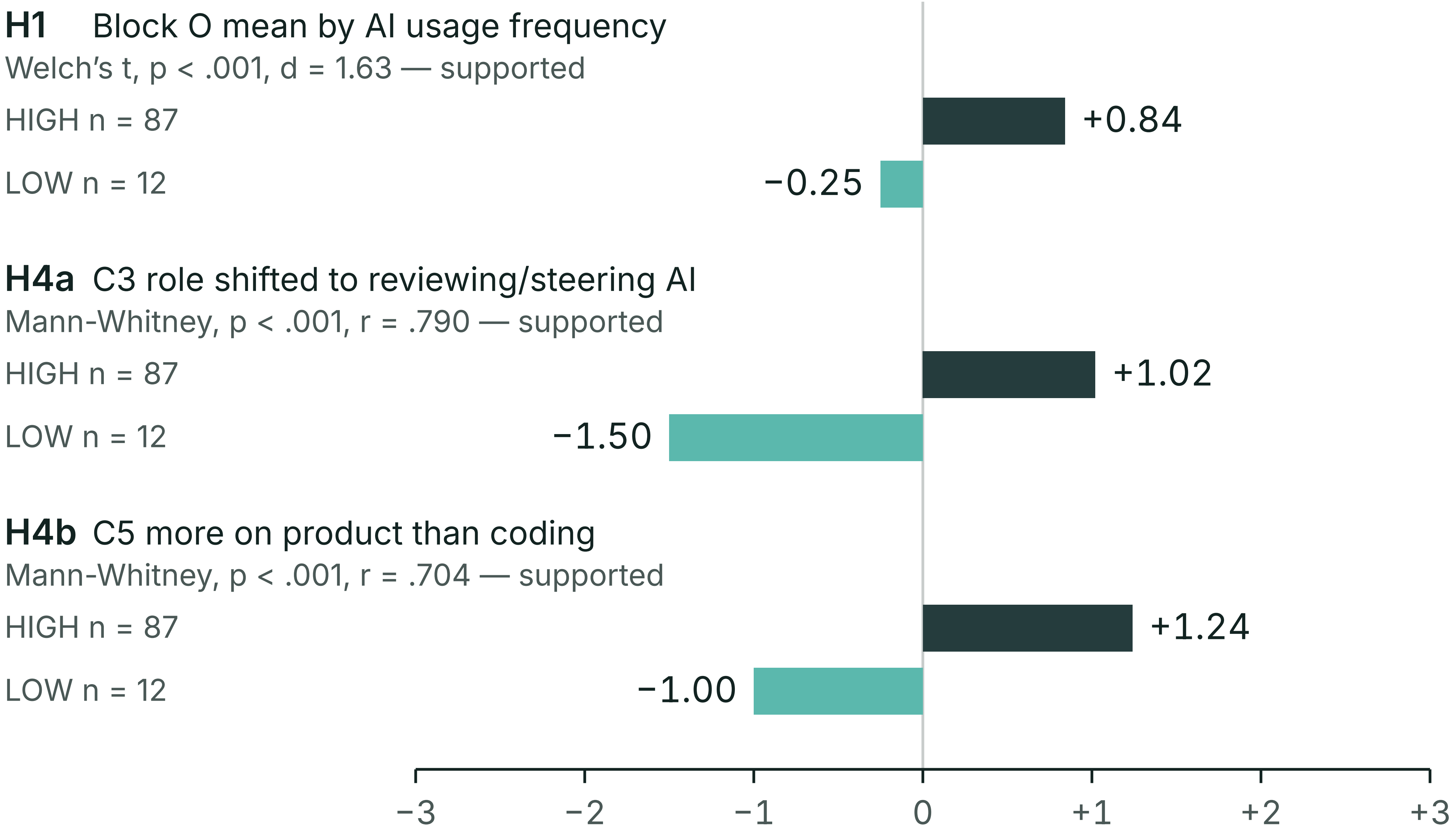


Fig. 5: H1, H4a and H4b by AI usage frequency. HIGH = daily or frequent use, n = 87; LOW = occasional or rare use, n = 12. Group means without CI — the paper reports none for the groups. This study, n = 99, scale −3 ... +3.

H1 to H4 · the four hypotheses from Section 1

H1 supported. The more intensively practitioners use AI, the less it looks like a tool problem.

H4a and H4b supported. The producer-to-steward shift is an **empirical reality** for high-frequency users, not a theoretical projection.

Hypothesis tests

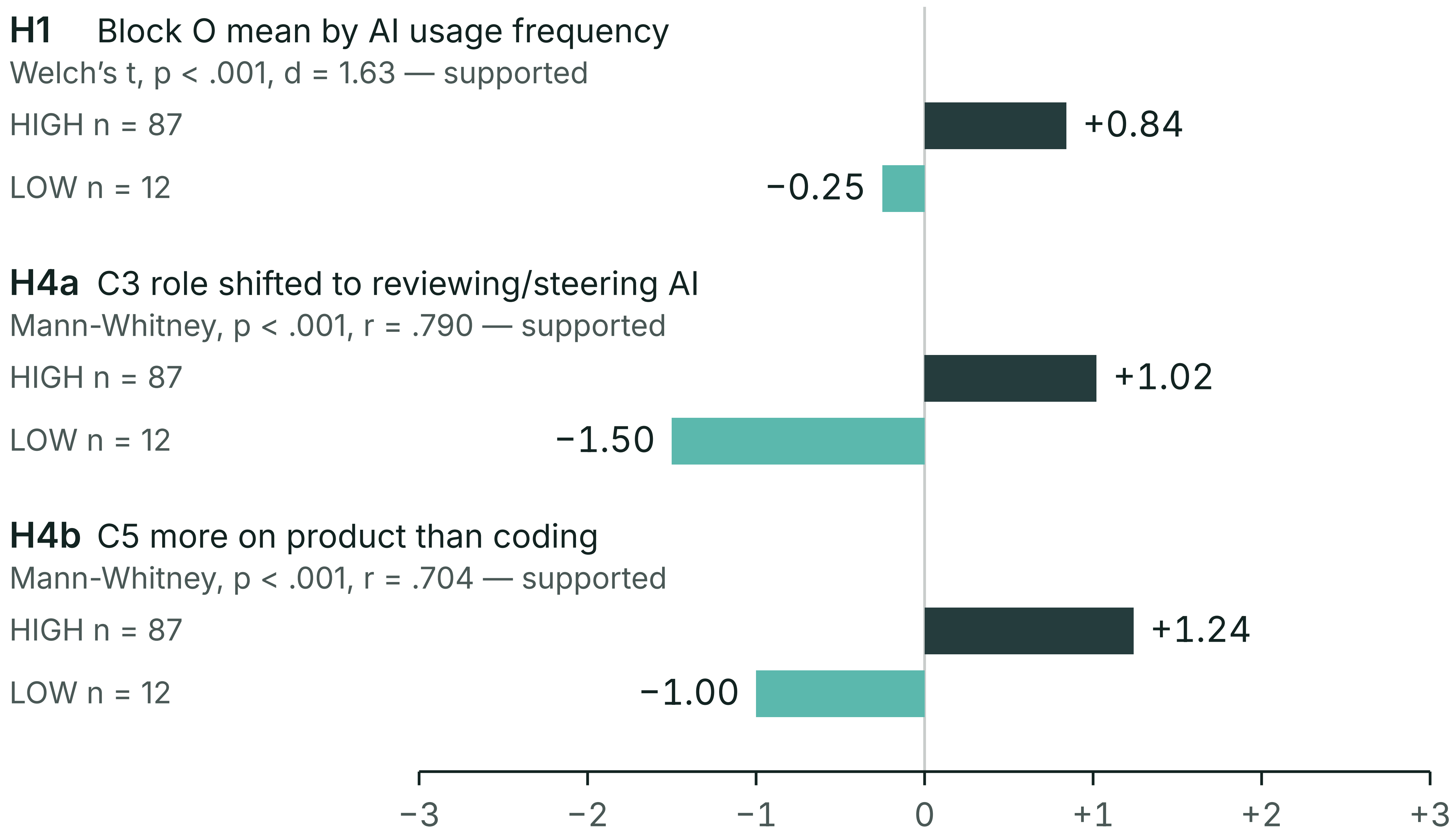


Fig. 5: H1, H4a and H4b by AI usage frequency. HIGH = daily or frequent use, n = 87; LOW = occasional or rare use, n = 12. Group means without CI — the paper reports none for the groups. This study, n = 99, scale −3 ... +3.

H1 to H4 · the four hypotheses from Section 1

H1 supported. The more intensively practitioners use AI, the less it looks like a tool problem.

H4a and H4b supported. The producer-to-steward shift is an **empirical reality** for high-frequency users, not a theoretical projection.

H3 · ρ = .381, $p < .001$, n = 99, supported. Perceived fragmentation correlates with the structural-need composite O1 + O3 + O5. Fragmentation functions as a leading indicator.

Hypothesis tests

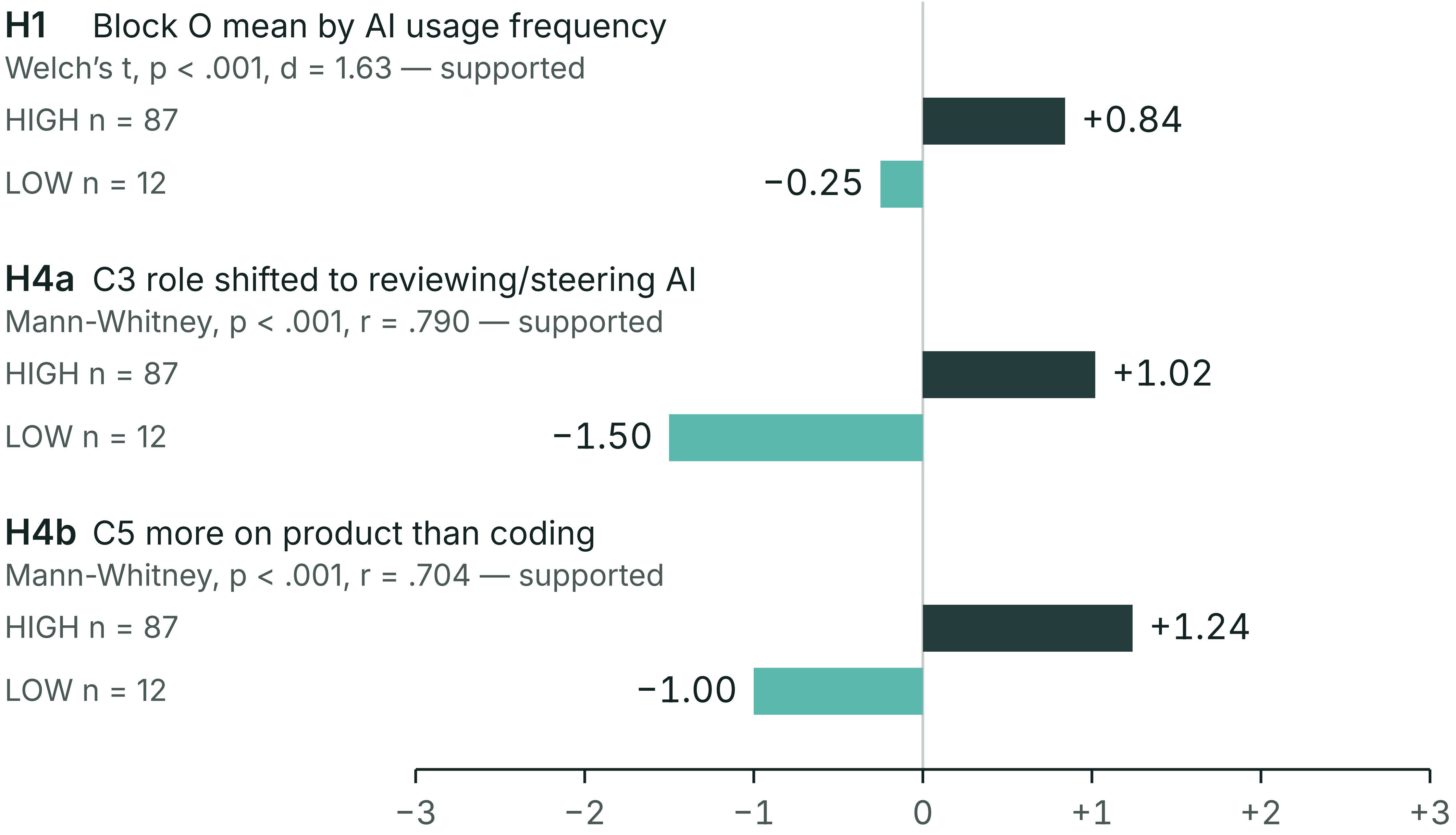


Fig. 5: H1, H4a and H4b by AI usage frequency. HIGH = daily or frequent use, n = 87; LOW = occasional or rare use, n = 12. Group means without CI — the paper reports none for the groups. This study, n = 99, scale −3 ... +3.

H1 to H4 · the four hypotheses from Section 1

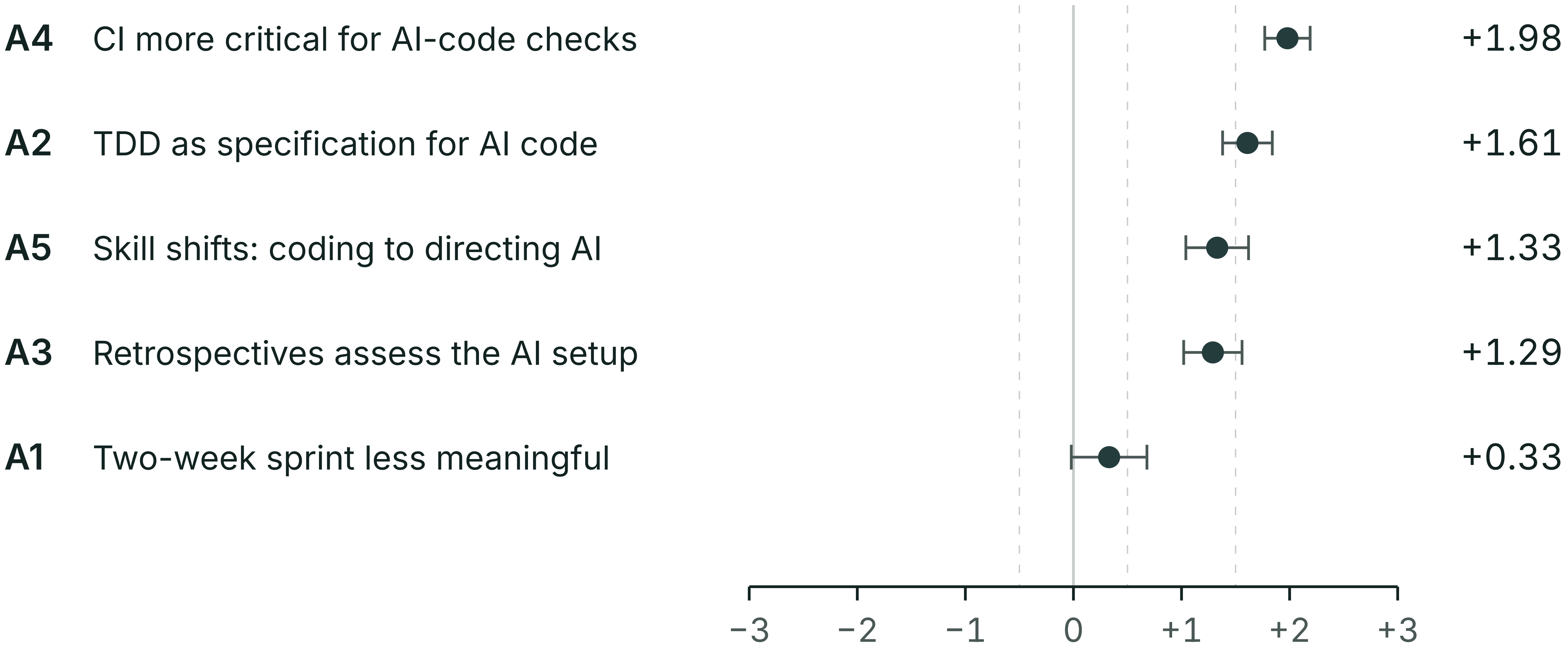
H1 supported. The more intensively practitioners use AI, the less it looks like a tool problem.

H4a and H4b supported. The producer-to-steward shift is an **empirical reality** for high-frequency users, not a theoretical projection.

H3 · ρ = .381, p < .001, n = 99, supported. Perceived fragmentation correlates with the structural-need composite O1 + O3 + O5. Fragmentation functions as a leading indicator.

H2 · not supported. SMALL n = 88, M = +1.12 vs. LARGE n = 11, M = +0.90; Mann-Whitney, p = .554, r = 0.11. Direction reversed, underpowered.

A governance envelope around AI code generation



Strong ≥ +1.5 · Moderate +0.5 to +1.5 · Neutral −0.5 to +0.5 · Disagreement < −0.5

Block A · RQ3 · descriptive, no hypothesis

The two strongest items in the whole survey both concern **quality governance**: Continuous Integration becomes more critical, and TDD is elevated from a coding discipline to a **specification mechanism for AI output.**

Selective transformation, not wholesale disruption: retrospectives expand to include the AI environment, core skills shift towards directing AI.

A1 is the only Neutral item and carries the highest dispersion in the survey, SD 1.77. The two-week sprint is neither strongly defended nor widely abandoned.

Fig. 6: Block A, agile practices in transition. Item means with 95 % confidence intervals; this study, n = 99, scale −3 … +3.

Open-ended responses: n = 58, five themes

What changes

- Review shifts from checking human code to **auditing AI output** for logical soundness and architectural fit → C3, H4a
- Continuous Integration and automated tests elevated to **safety infrastructure** → A4, A2
- Planning and estimation **disrupted** as AI collapses multi-day tasks

What holds

- The two-week sprint **adapts under pressure** rather than being abandoned → A1 and its dispersion
- A recurring minority reports **no change,** treating AI as a personal productivity layer on unchanged ceremonies → the low-frequency users in H4

Section 4

Implications

Organisations have a window of opportunity to respond structurally before fragmentation becomes entrenched.

Hinderks, Thomaschewski & Schön · ISD 2026

What follows

- Treat AI development environments as **shared infrastructure maintained by dedicated teams,** not assembled ad hoc.
- For small teams, the **simplified version of the model** is the realistic entry point.
- **CI and TDD are the governance envelope** — with a direct addressee, the Platform Team layer.
- **Monitor fragmentation as the leading indicator,** before it becomes a complaint.

Limitations

Measurement

- The blocks are **thematic item groups,** not validated reflective scales; targeted sub-scales still have to be developed and validated.
- **Perceptions may diverge from organisational reality** — the plausibility-to-applicability gap is itself evidence of this.

Sample and design

- **Opt-in Prolific sample,** likely over-representing tech-savvy professionals; 87.9 % daily AI users.
- **Cross-sectional design** cannot establish causality.
- A **March 2026 snapshot.** Given how fast AI coding tools evolve, the findings are time-indexed.

Future work

- **Longitudinal studies** tracking structural evolution as adoption deepens.
- **Qualitative interviews** to explore the plausibility-to-applicability gap.
- **Validated sub-scales** for the multi-dimensional constructs in Blocks O and M.
- **Case studies** in organisations already running dedicated AI environment teams.

Practitioners endorse shared infrastructure, find the model plausible but not yet applicable, and — where AI use is intensive — already live the shift from producing code to steering it.

References

Herda, T., Pichler, V., Zhang, Z., Abrahamsson, P., & Hanssen, G. K. (2025). AI and agile software development: From frustration to success – XP2025 workshop summary. *Lecture Notes in Business Information Processing, 561,* 3–13. https://doi.org/10.1007/978-3-032-05799-0_1

Hinderks, A., Thomaschewski, J., & Schön, E.-M. (2026). From AI assistants to AI-Dev-Systems: A three-layer model for human–AI collaboration in agile software development. *Lecture Notes in Business Information Processing.* Springer. (Accepted; to appear)

Kudina, O., & van de Poel, I. (2024). A sociotechnical system perspective on AI. *Minds and Machines, 34*(21), 1–9. https://doi.org/10.1007/s11023-024-09676-8

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). *The impact of AI on developer productivity: Evidence from GitHub Copilot.* arXiv:2302.06590.

References (continued)

Schön, E.-M., Silva da Silva, T., Hinderks, A., Sharp, H., & Thomaschewski, J. (2023). Introduction to special issue on Agile UX: Challenges, successes and barriers to improvement. *Information and Software Technology, 158,* 107193. https://doi.org/10.1016/j.infsof.2023.107193

Skelton, M., & Pais, M. (2019). *Team topologies: Organizing business and technology teams for fast flow.* IT Revolution Press.

Wivestad, V. T., Barbala, A., & Stray, V. (2024). Copilot's island of joy: Balancing individual satisfaction with team interaction in agile development. *Lecture Notes in Business Information Processing,* 123–129. https://doi.org/10.1007/978-3-031-61309-9_8

Ziegler, A., Kalliamvakou, E., Li, X. A., Rice, A., Rifkin, D., Simister, S., Sittampalam, G., & Aftandilian, E. (2024). Measuring GitHub Copilot's impact on productivity. *Communications of the ACM, 67*(3), 54–63. https://doi.org/10.1145/3633453

Thank you

Andreas Hinderks

jan-andreas.hinderks@hs-hannover.de · hinderks.org

Jörg Thomaschewski · Eva-Maria Schön, University of Applied Sciences Emden/Leer

Backup · Table 2, Block C

Item	Construct (paraphrased)	M	SD	95 % CI	Interpretation
C1	AI changed how I approach coding	+1.98	1.18	[+1.74, +2.22]	Strong Agr.
C2	Coding is no longer the bottleneck	+0.94	1.56	[+0.63, +1.25]	Moderate Agr.
C3	Role shifted to reviewing/steering AI	+0.72	1.69	[+0.38, +1.06]	Moderate Agr.
C4	Team iterates much faster than before	+1.53	1.45	[+1.24, +1.82]	Strong Agr.
C5	More on product than coding	+0.97	1.59	[+0.65, +1.29]	Moderate Agr.
C6	Different AI tools create friction	−0.66	1.53	[−0.97, −0.35]	Disagreement
C7	Team lacks clear AI usage guidelines	−0.22	1.61	[−0.54, +0.10]	Neutral
C8	AI pair programming feels different	+1.47	1.26	[+1.22, +1.72]	Moderate Agr.

Block mean +0.84

This study, n = 99, scale −3 … +3.

Backup · Table 2, Block O

Item	Construct (paraphrased)	M	SD	95 % CI	Interpretation
O1	Team restructuring, not just tools	+0.70	1.57	[+0.39, +1.01]	Moderate Agr.
O2	“Copilot” framing downplays impact	+0.74	1.60	[+0.42, +1.06]	Moderate Agr.
O3	AI platforms need dedicated teams	+1.13	1.33	[+0.86, +1.40]	Moderate Agr.
O4	Independent tool choice fragments	+0.46	1.62	[+0.14, +0.78]	Neutral
O5	AI tools as shared infrastructure	+1.38	1.27	[+1.13, +1.63]	Moderate Agr.
O6	AI-code governance is insufficient	+0.62	1.54	[+0.31, +0.93]	Moderate Agr.
O7	Technical, not organisational ®	+0.08	1.57	[−0.23, +0.39]	Neutral

Block mean +0.73

® reverse-coded item.

This study, n = 99, scale −3 … +3.

Backup · Table 3, Block M

Item	Construct (paraphrased)	M	SD	95 % CI	Interpretation
M1	Three-layer model is plausible	+1.49	1.10	[+1.27, +1.71]	Moderate Agr.
M2	System Teams ease choice overload	+1.39	1.22	[+1.15, +1.63]	Moderate Agr.
M3	Platform Teams improve consistency	+1.27	1.33	[+1.00, +1.54]	Moderate Agr.
M4	Prefer curation over self-selection	+0.80	1.70	[+0.46, +1.14]	Moderate Agr.
M5	Model fits my organisation	+0.76	1.53	[+0.45, +1.07]	Moderate Agr.
M6	Needs org. maturity most firms lack	+1.40	1.38	[+1.12, +1.68]	Moderate Agr.
M7	Simplified version fits small teams	+1.12	1.55	[+0.81, +1.43]	Moderate Agr.
M8	Model risks new silos between teams	+0.52	1.38	[+0.24, +0.80]	Neutral

Block mean +1.09

M8 carries the only negative item-total correlation, $r = -.414$.

This study, $n = 99$, scale −3 ... +3.

Backup · Table 3, Block A

Item	Construct (paraphrased)	M	SD	95 % CI	Interpretation
A1	Two-week sprint less meaningful	+0.33	1.77	[−0.02, +0.68]	Neutral
A2	TDD as specification for AI code	+1.61	1.14	[+1.38, +1.84]	Strong Agr.
A3	Retrospectives assess the AI setup	+1.29	1.35	[+1.02, +1.56]	Moderate Agr.
A4	CI more critical for AI-code checks	+1.98	1.07	[+1.77, +2.19]	Strong Agr.
A5	Skill shifts: coding to directing AI	+1.33	1.47	[+1.04, +1.62]	Moderate Agr.

Block mean +1.31

A1 carries the highest dispersion in the survey, SD 1.77.

This study, n = 99, scale −3 … +3.

Backup · Table 4, hypothesis tests

H	Test	Groups / descriptives	p	Effect	Verdict
H1	Welch's t	HIGH ($n = 87$, $M = +0.84$) vs. LOW ($n = 12$, $M = -0.25$)	$< .001$	$d = 1.63$	Supported
H2	Mann-Whitney	SMALL ($n = 88$, $M = +1.12$) vs. LARGE ($n = 11$, $M = +0.90$)	.554	$r = 0.11$	Not supported
H3	Spearman ρ	C6 ↔ O1 + O3 + O5 ($n = 99$)	$< .001$	$\rho = .381$	Supported
H4a	Mann-Whitney	HIGH ($M = +1.02$) vs. LOW ($M = -1.50$)	$< .001$	$r = .790$	Supported
H4b	Mann-Whitney	HIGH ($M = +1.24$) vs. LOW ($M = -1.00$)	$< .001$	$r = .704$	Supported

This study, $n = 99$.

Bonferroni-corrected for multiple comparisons.

Backup · Table 1, instrument overview

Block	Construct	Items	RQ / H	Example item (paraphrased)
C	Perceived changes from AI tools	8	RQ3, H4	AI tools have fundamentally changed how I approach coding tasks
O	AI integration as organisational challenge	7	RQ1, H1, H3	AI integration requires changes to team structures, not just tool adoption
M	Three-layer model plausibility	8	RQ2, H2	The three-layer model is a plausible way to organise AI-augmented development
A	Agile practices in transition	5	RQ3	TDD is more important as a specification mechanism for AI-generated code

28 items, 7-point scale normalised to −3 ... +3, plus one open-text question.

Full verbatim instrument in the online appendix.