

Supplementary Information

Orchestrating human–AI software delivery: a retrospective longitudinal field study of three software modernization programs

Maximiliano Armesto^{1,*} and Christophe Kolb¹

¹ Taller Technologies, Córdoba, Argentina.

*Corresponding author: maximiliano.armesto@tallertechnologies.com

Contents

- Supplementary Table S1. Two informative contrasts for the mature V4 configuration.
- Supplementary Table S2. Banking application benchmark across delivery configurations.
- Supplementary Table S3. ACAS accounting platform benchmark across delivery configurations.
- Supplementary Table S4. Mortgage application benchmark across delivery configurations.
- Supplementary Table S5. Modelled effort and observed validation-stage issue load by program and configuration.
- Supplementary Table S6. Metric taxonomy used in the study.
- Supplementary Table S7. Leave-one-program-out sensitivity for the V4 versus traditional comparison.
- Supplementary Table S8. Approximate V4 review-stage containment.
- Supplementary Note 1. Reproducing the reported aggregates.
- Supplementary Data 1. Complete project-version dataset (supplied separately as an .xlsx file).

Table S1. Two informative contrasts for the mature V4 configuration. The V1-to-V4 contrast is especially useful because both configurations share the same nominal agentic staffing scenario (five people, three senior equivalents), so the reduction is not driven by a change in staffing composition. Time reduction, raw-effort reduction, SEE reduction and issue-load reduction are percentage reductions; coverage change is in percentage points (pp).

Contrast	Speed-up	Time red.	Raw-effort red.	SEE red.	Issue-load red.	Coverage change
Traditional $\rightarrow$ V4	3.87 $\times$	74.2%	78.5%	87.1%	74.0%	+13.4 pp
V1 $\rightarrow$ V4	3.08 $\times$	67.5%	67.5%	67.5%	75.8%	+37.9 pp

Table S2. Banking application benchmark across delivery configurations. Stage durations are in weeks. Tasks are recorded backlog items; issues are those reaching downstream validation; coverage is the percentage of first-release requirements completed and accepted at initial handoff.

Configuration	Analysis	Planning	Impl.	Valid.	Total	Tasks	Issues	Coverage
Traditional	2.0	2.0	4.0	2.0	10.0	400	35	80%
V1	1.0	0.4	3.0	3.0	7.4	800	80	55%
V2	0.8	0.4	2.5	3.0	6.7	800	76	60%
V3	0.6	0.3	1.5	1.5	3.9	800	40	85%
V4	0.2	0.2	1.0	1.0	2.4	800	20	90%

Impl., implementation; Valid., validation.

Table S3. ACAS accounting platform benchmark across delivery configurations. Columns are defined as in Supplementary Table S2.

Configuration	Analysis	Planning	Impl.	Valid.	Total	Tasks	Issues	Coverage
Traditional	5.0	3.0	8.0	4.0	20.0	2500	200	75%
V1	2.5	1.0	6.0	7.0	16.5	4500	380	50%
V2	2.2	0.9	5.2	7.0	15.3	4500	360	55%
V3	1.5	0.6	3.2	2.0	7.3	4500	180	82%
V4	0.8	0.6	2.0	1.5	4.9	4500	90	90%

Impl., implementation; Valid., validation.

Table S4. Mortgage application benchmark across delivery configurations. Columns are defined as in Supplementary Table S2.

Configuration	Analysis	Planning	Impl.	Valid.	Total	Tasks	Issues	Coverage
Traditional	1.5	1.0	2.0	1.5	6.0	300	22	90%
V1	0.8	0.4	1.5	2.0	4.7	550	45	70%
V2	0.7	0.4	1.3	2.0	4.4	550	42	75%
V3	0.4	0.2	1.0	1.0	2.6	550	25	92%
V4	0.2	0.2	0.8	0.8	2.0	550	12	95%

Impl., implementation; Valid., validation.

Table S5. Modelled effort and observed validation-stage issue load by program and configuration. Raw person-days and SEE-days are modelled under the staffing scenarios defined in the Methods and are not direct labour-hour logs. Issue load is observed: issues reaching downstream validation per 100 recorded tasks.

Program	Configuration	Raw person-days	SEE-days	Issue load per 100 tasks
Banking	Traditional	300.0	300.0	8.75
Banking	V1	185.0	111.0	10.00
Banking	V2	167.5	100.5	9.50
Banking	V3	97.5	58.5	5.00
Banking	V4	60.0	36.0	2.50
ACAS	Traditional	600.0	600.0	8.00
ACAS	V1	412.5	247.5	8.44
ACAS	V2	382.5	229.5	8.00
ACAS	V3	182.5	109.5	4.00
ACAS	V4	122.5	73.5	2.00
Mortgage	Traditional	180.0	180.0	7.33
Mortgage	V1	117.5	70.5	8.18
Mortgage	V2	110.0	66.0	7.64
Mortgage	V3	65.0	39.0	4.55
Mortgage	V4	50.0	30.0	2.18

SEE, senior equivalent.

Table S6. Metric taxonomy used in the study. Observed metrics are recorded delivery outcomes; modelled metrics are scenario-based interpretations of those outcomes.

Metric	Type	Definition	Interpretation caveat
Total duration $T_{p,v}$	Observed	Sum of measured analysis, planning, implementation and validation durations.	Excludes deployment, because that stage was not represented consistently across configurations.
Task count $N_{p,v}$	Observed	Number of recorded work items in the program backlog for the project-version cell.	Task granularity changes materially across delivery configurations.
Validation issues $I_{p,v}$	Observed	Issues reaching downstream validation.	Captures defect escape to validation, not total defects created anywhere in the pipeline.
Validation-stage issue load $L_{p,v}$	Derived from observed measures	$100 \times I_{p,v}/N_{p,v}$, reported as issues per 100 recorded tasks.	A workload-normalized escape metric; not equivalent to intrinsic defect density, because the denominator is intervention-sensitive.
First-release coverage $C_{p,v}$	Observed	Fraction of first-release requirements recorded as completed and accepted at initial handoff.	Sensitive to requirement scoping and adjudication practices.
Raw person-days $E_{p,v}^{\text{raw}}$	Modelled	$5h_v T_{p,v}$ under full-time staffing with $h_{\text{Traditional}} = 6$ and $h_{\text{Agentic}} = 5$.	A scenario-based effort estimate, not directly logged labour-hours.
SEE-days $E_{p,v}^{\text{SEE}}$	Modelled	$5\sigma_v T_{p,v}$ with $\sigma_{\text{Traditional}} = 6$ and $\sigma_{\text{Agentic}} = 1 + 4w = 3$ at the baseline junior weight $w = 0.5$.	Useful for staffing scenarios, but dependent on the chosen junior-to-senior weighting.
Review containment R_p	Derived from observed measures	$1 - I_{p,V4}/I_p^{\text{obs}}$, where I_p^{obs} is all issues observed across review and validation.	Available for the V4 configuration only; counts are approximate.

Table S7. Leave-one-program-out sensitivity for the V4 versus traditional portfolio comparison. The duration ratio is the summed V4 duration divided by the summed traditional duration for the retained programs; the issue-load ratio is the same ratio computed for validation-stage issue load. Coverage change is in percentage points (pp).

Subset	Duration ratio	Issue-load ratio	Coverage change
All programs	0.258	0.260	+13.4 pp
Excluding banking	0.265	0.255	+13.9 pp
Excluding ACAS	0.275	0.291	+7.8 pp
Excluding mortgage	0.243	0.256	+14.3 pp

Table S8. Approximate V4 review-stage containment. Total observed issues are all issues recorded in the V4 configuration across review and validation; the issues reaching validation are a subset of that total. The containment rate is one minus the ratio of validation issues to total observed issues, as in equation (4) of the main text. Counts are approximate and were available for the V4 configuration only.

Program	Total observed issues	Issues reaching validation	Containment rate
Banking	40	20	50.0%
ACAS	188	90	52.1%
Mortgage	23	12	47.8%
Portfolio (weighted)	251	122	51.4%

Supplementary Note 1: Reproducing the reported aggregates

Every aggregate reported in the main text can be recomputed from Supplementary Tables S2–S4 (equivalently, from Supplementary Data 1) using the definitions given in the Methods. This note records the arithmetic so that the derivation chain is fully auditable.

Summed portfolio duration. $\sum_p T_{p,v}$ over the three programs gives 36.0, 28.6, 26.4, 13.8 and 9.3 weeks for Traditional, V1, V2, V3 and V4 respectively.

Modelled effort. Raw person-days are $5h_v \sum_p T_{p,v}$ with $h_{\text{Traditional}} = 6$ and $h_{\text{Agentic}} = 5$, giving 1080.0, 715.0, 660.0, 345.0 and 232.5. SEE-days are $5\sigma_v \sum_p T_{p,v}$ with $\sigma_{\text{Traditional}} = 6$ and $\sigma_{\text{Agentic}} = 3$, giving 1080.0, 429.0, 396.0, 207.0 and 139.5.

Portfolio issue load. $100 \sum_p I_{p,v} / \sum_p N_{p,v}$ gives $100 \times 257/3200 = 8.03$, $505/5850 = 8.63$, $478/5850 = 8.17$, $245/5850 = 4.19$ and $122/5850 = 2.09$ issues per 100 recorded tasks. The denominator changes between the traditional baseline (3200 tasks) and the agentic configurations (5850 tasks) because agentic configurations produce more granular backlogs; this is the denominator-drift caveat discussed in the Discussion.

Task-weighted coverage. $\bar{C}_v = \sum_p C_{p,v} N_{p,v} / \sum_p N_{p,v}$ gives 77.0%, 52.6%, 57.6%, 83.4% and 90.5%.

Headline contrasts. Traditional to V4: speed-up $36.0/9.3 = 3.87$; time reduction $1 - 9.3/36.0 = 74.2\%$; raw-effort reduction $1 - 232.5/1080.0 = 78.5\%$; SEE reduction $1 - 139.5/1080.0 = 87.1\%$; issue-load reduction $1 - 2.09/8.03 = 74.0\%$; coverage change $90.5 - 77.0 = +13.4$ pp. V1 to V4: speed-up $28.6/9.3 = 3.08$; time and modelled-effort reduction $1 - 9.3/28.6 = 67.5\%$ (identical for both effort measures because V1 and V4 share the same staffing scenario); issue-load reduction $1 - 2.09/8.63 = 75.8\%$; coverage change $90.5 - 52.6 = +37.9$ pp.

Senior-equivalent sensitivity. With each junior member weighted at w senior equivalents, agentic staffing is $1 + 4w$ and V4 SEE-days are $5(1 + 4w) \times 9.3$, giving 93.0, 139.5, 186.0 and 232.5 for $w = 0.25, 0.50, 0.75, 1.00$ (Table 4 of the main text).

Review containment. Containment follows equation (4) of the main text, $R_p = 1 - I_{p,V4}/I_p^{\text{obs}}$, giving $1 - 20/40 = 50.0\%$, $1 - 90/188 = 52.1\%$ and $1 - 12/23 = 47.8\%$; the portfolio-weighted rate is $1 - 122/251 = 51.4\%$.

Supplementary Data 1

Supplementary Data 1 is supplied separately as an `.xlsx` file containing the complete dataset used in this study, across four sheets.

`project_version_dataset` holds one row per project-version cell (3 programs $\times$ 5 delivery configurations = 15 rows), with columns for program, configuration, the four stage durations, total duration, task count, issues reaching validation, first-release coverage, and the derived issue-load and modelled-effort quantities. Every column is populated for every row.

`v4_review_containment` holds the review-stage containment counts. These were recorded for the V4 configuration only, so they are kept in a separate 3-row sheet rather than left as empty cells in

the main dataset. Its `issues_reaching_validation` column repeats the V4 rows of the main sheet, and `total_observed_issues` is the larger figure of which those issues are a subset.

`data_dictionary` defines every column in both data sheets, and `notes` records the provenance and the observed-versus-modelled distinction. Comma-separated copies of both data sheets are provided for convenience.