Akholi

PART 3 / 2026 RESEARCH STUDY

Governance, Oversight, and Control

Agentic AI Outsourcing at Tier 1 Financial Institutions

Phil Hatch

150 engagements · 850 delivery-team respondents

Observational, cross-sectional study of a nonrepresentative sample. Financial institutions did not independently validate the reported experience. The findings do not establish causation or legal breach.

Version 1.8.2 | November 1, 2026

REPORT INFORMATION

Agentic AI Outsourcing at Tier 1 Financial Institutions

Full Study Report

Author	Phil Hatch
Publication date	November 1, 2026
Version	1.8.2
ISBN	979-8-234-26882-2
SSRN DOI	https://doi.org/10.2139/ssrn.7164559
Figshare DOI	https://doi.org/10.6084/m9.figshare.34036890
Contact	info@akholi.com
Website	akholi.com

AUTHOR ONLINE

Akholi author page | ORCID | LinkedIn | ResearchGate

SUGGESTED CITATION

Hatch, Phil. 2026. *Agentic AI Outsourcing at Tier 1 Financial Institutions: Full Study Report*. Version 1.8.2. Akholi. https://doi.org/10.2139/ssrn.7164559.

COPYRIGHT

© 2026 Phil Hatch and Akholi. All rights reserved.

No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the copyright holders, except for brief quotations with full attribution in reviews, research, and other noncommercial uses permitted by copyright law. Requests for permission should be sent to info@akholi.com.

USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence tools were used to prepare this report. Appendix G discloses how they were used.

Contents

Governance, Oversight, and Control **5**
Reported Governance Weaknesses Across All 150 Engagements **7**
1 The Governance, Oversight, and Control Score 13
2 Governance Indicators Across the Five Failure Categories 23
3 Governance Across the Segments of the Study 35
4 Governance by Lifecycle Stage, Engagement Age, and Expected Cancellation 49
5 Governance Against the Four Study Measures 55

Co-Management and Governance Across the Other Four Categories **63**
1 Joint Governance and Decision Rights 67
2 Shared Risk and Issue Management 77
3 Joint Change Control 87
4 Human Review and the Joint Quality Assurance Loop 95
5 The Shift From Managing People to Managing Agents 107

Financial Institution Readiness and the Governance of Agentic Work **115**
1 Financial-Institution-Provided Data 117
2 Management of the Financial Institution's Side of the Engagement 121
3 Governance of Agentic AI Within the Financial Institution 127
4 Provisioning of System Access and Agent Identities 133

Vendor Preparation and Accountability in the Control Environment **137**
1 Delivery Playbook, Quality Assurance, and Platforms 139
2 Accountability for Agent Error 143

Technical Controls and Governance Across the Engagement **147**
1 Data Isolation, Sensitive Data, and Adversarial Input 149
2 Authorized Action and Guardrails 153
3 Testability and Behavioral Stability 159
4 Observability of Agent Work 161

Third-Party Dependencies and the Limits of Internal Control **169**
1 Notice of Provider Change 173
2 Detection of Unannounced Change 177
3 Contractual and Licensing Coverage of Agent Work 185
4 Security Posture and Provider Concentration 189
5 Fallback Capacity and Retention of Operational Knowledge 193
6 The Dependency Chain and Gated External Rails 197

Understanding Governing Obligations and the Control Environment **201**
1 Knowledge of the Governing Obligations 205
2 Discussion and Training on the Governing Obligations 213

The Fit Between Governance Processes and the Pace of Agent Work **217**
1 Slow Response From the Vendor and the Financial Institution 221
2 Governance Functions Inherited From Before Agentic AI 229
3 The Newness of Agentic AI to Both Firms 243
4 Volume and Speed of AI Output Beyond the Controls 251
5 Non-Deterministic Output and Missing Logs 255
6 Respondent Explanations Locating Problems Upstream 261

Manual Alteration and the Operation and Visibility of Controls **263**
1 The Five Manual Alteration Methods and the Score 265
2 Oversight of Reporting Provided to the Financial Institution 281
3 Escalation, Disclosure, and Record of Manual Alteration 305
4 Additional Reporting Practices Below the Threshold for Statistical Analysis 333

Practices Adopted in Place of Controls **339**
1 Workarounds Adopted as Standard Practice 341
2 Remedies Proposed by Respondents 351

About the Author **353**
About Akholi **355**

Governance, Oversight, and Control

The delivery workforce reported governance and control weaknesses across all 150 engagements. The central problem was the operation of controls in shared agentic work: authority, change management, human review, and evidence did not consistently provide the assurance their formal arrangements implied. The governance, oversight, and control score combines 47 related indicator ratings. Lower scores were associated with cancellation, manual correction, and limited visibility of that work. The mean score was 2.58 on canceled engagements and 4.86 on active engagements. These findings connect operational control with engagement performance, while leaving causal direction unresolved.

Part 3 examines how indicators, engagement characteristics, and reported practices relate to governance, oversight, and control. It draws on 37,472 ratings of 47 governance-related indicators from 850 respondents, together with reporting and obligations items and written accounts. The 3,330 written instances classified as governance weaknesses occur among 6,862 answers concerning respondents' own engagements. They identify reported deficiencies and bypassed controls, not independently verified failure of every engagement. The analysis considers both broad patterns in the score and specific control gaps that a composite mean may not sharply distinguish.

The evidence principally records the vendor delivery workforce's experience at one observation. It cannot establish the sequence of events or identify causal effects. The 56 canceled engagements had already been canceled when respondents answered, making comparisons with the 94 active engagements retrospective. Two customer management executives are included among the 850 respondents, but they do not provide a corresponding evidence base from financial institution staff. Relationships among measures from the same respondents also share that source limitation.

“

What we did was stop using the change board for anything that mattered. It met fortnightly and the agents moved faster than that, so we built a parallel list, one sheet, updated every evening, of every case where the model had done something we had not agreed to and what we had done about it. Their underwriting lead and I read it at nine each morning and cleared it between us, and whatever survived that went to the formal forum a fortnight later as a note rather than a decision. It worked, in the sense that files kept moving. It also means the actual governance of this engagement was two people and a spreadsheet. Nobody asked us to do it and nobody stopped us, which I think tells you where the accountability sat.”

ENGAGEMENT MANAGER

BPO, Credit, underwriting and suitability

Reported Governance Weaknesses Across All 150 Engagements

Reported control weaknesses were widespread, but their depth varied. Ratings of 3 or below occurred on 145 of the 150 engagements, while written accounts classified as governance weaknesses occurred on all 150. Only 50 engagements had a mean governance, oversight, and control score of 3 or below. These are different findings: the presence of one reported weakness is not equivalent to a low composite mean or independently established overall governance failure. The canceled-versus-active contrast was substantial, whereas structural group comparisons were less conclusive.

The governance, oversight, and control score is the mean of the 47 governance-related indicator ratings available for each respondent. Respondent means are averaged within engagements, and engagement means are then averaged across the study. Low-rating counts report ratings of 3 or below separately from the score. Written instances count answers classified negative on a governance indicator or describing one of the 26 forms in Table 3.1. These measures describe different aspects of reported control weakness. Numeric score bands are descriptive and do not represent validated stages of control development.

The measures were strongly related across the 150 engagements. Correlation, denoted r, expresses the direction and strength of a linear association on a scale from −1 to +1. The score correlated at −0.97 with low ratings per respondent and at −0.78 with written instances per respondent. These are complementary descriptions from shared respondents and source material, rather than independent validation. The five engagements with no low governance rating were active and in the highest indicator quintile, the highest fifth ranked by the mean of all 100 engagement indicators. Their scores ranged from 6.62 to 7.38, yet each had 5–15 written instances of weakness. The median engagement had 21.50 such instances, within a range of 5–44.

Table 3.1. The 26 forms of a control bypassed or not working in the written answers

Form	Definition
Financial institution explicitly not informed	The respondent stated that the financial institution was not told of the manual alteration.
Manual alteration disclosed to no party	The manual alteration was disclosed to no party.
Manual alteration not logged or only partly logged	The manual alteration did not enter a record, or only part of it did.

Table 3.1. The 26 forms of a control bypassed or not working in the written answers (continued)

Form	Definition
Instructed not to raise the problem	The respondent was told to make the work succeed without escalating, or that raising the problem would be penalized.
Raising problems seen as futile	The respondent had stopped raising the problem, since nothing changed or raising it cost more than correcting it.
Problems raised with no response	The respondent raised the problem and received no response.
Effort absent from reported figures	The work performed appeared in no reported measure.
Human work presented as agentic	A person performed work that was reported or structured as work performed by an agent.
Habit in place of a control	The compensating practice was undocumented and depended on one person.
Failures removed from the base of a reported rate	Failures or exceptions were excluded from the denominator of a reported rate.
Measures restricted to the parseable subset	A measure covered only the easy or machine-readable part of the work.
Human-corrected work counted as agent work	Work corrected by people, run in a test environment, or overridden was counted as production work completed by agents.
Accuracy reported on a selected validation set	Accuracy was reported on a set chosen to exclude the hardest cases.
Elapsed-time measures restarted on rework	The measurement of elapsed time restarted on rework, so delay never accumulated.
Activity counted in place of outcome	A count of activity stood in for a measure of outcome.
Exceptions omitted from reports	An exception recorded internally was omitted from what the financial institution received.
Readiness or completion reported falsely	A readiness or completion status was reported falsely, allowing a gate to pass.
Suppressed aging of open items	Open items were pooled into a category that did not age, so their duration never appeared.
Schema or format changed without notice	A layout, field name, or column order changed between deliveries without notice.
No validation by the financial institution before handover	The financial institution performed no check before handing data or documents over.

Table 3.1. The 26 forms of a control bypassed or not working in the written answers (continued)

Form	Definition
Undefined handoffs between the financial institution's teams	Responsibility between the financial institution's own teams was not defined.
Output that cannot be signed off	Output that did not reproduce could not be attributed or approved.
Unclear liability or regulatory accountability	The respondent could not say who was legally or regulatorily answerable for the work or for an agent's error.
Regulatory or compliance exposure	The respondent linked a practice, gap, or incident to a breach, or the risk of a breach, of a legal, regulatory, or compliance obligation.
Change discovered only through failure	A change or outage was discovered when something downstream failed, never through notice.
No escalation path outside working hours	No route existed for an agent failure outside normal working hours.

Form	Written answers	Respondents	Engagements
Financial institution explicitly not informed	178	139	73
Manual alteration disclosed to no party	59	55	39
Manual alteration not logged or only partly logged	143	118	72
Instructed not to raise the problem	20	15	12
Raising problems seen as futile	149	122	81
Problems raised with no response	326	255	119
Effort absent from reported figures	71	64	54
Human work presented as agentic	6	6	5
Habit in place of a control	91	87	69
Failures removed from the base of a reported rate	50	50	39
Measures restricted to the parseable subset	38	38	32
Human-corrected work counted as agent work	52	52	45
Accuracy reported on a selected validation set	18	18	16

Table 3.1. The 26 forms of a control bypassed or not working in the written answers (continued)

Form	Definition			
Elapsed-time measures restarted on rework		4	4	3
Activity counted in place of outcome		112	111	78
Exceptions omitted from reports		23	23	20
Readiness or completion reported falsely		20	20	14
Suppressed aging of open items		3	3	3
Schema or format changed without notice		61	60	51
No validation by the financial institution before handover		50	47	37
Undefined handoffs between the financial institution's teams		41	39	35
Output that cannot be signed off		29	27	23
Unclear liability or regulatory accountability		83	80	57
Regulatory or compliance exposure		16	14	12
Change discovered only through failure		157	150	93
No escalation path outside working hours		14	14	14

Note: A written answer may describe more than one form. The forms are counted on the written answers that addressed the respondent's own engagement.

Respondent role, engagement status, expected cancellation, and the four study measures showed relationships with the reported control environment. The structural group comparisons did not detect adjusted differences on the stated measures. Team size and age use continuous analysis and an equivalence margin, the largest association treated as sufficiently small under the study's criterion. Table 3.2 summarizes those results. Nonsignificance does not establish identical governance, and equivalence concerns only the specified range of associations.

Table 3.2. Segments and study measures against the three measures of reported governance weakness

Segment or study measure	Unit	Score	Distressed governance ratings	Written instances
Financial institution type	engagement	No detected association (Holm p = 1.00)	No detected association (Holm p = 1.00)	No detected association (Holm p = 1.00)
Outsourcing model	engagement	No detected association (Holm p = 1.00)	No detected association (Holm p = 1.00)	No detected association (Holm p = 1.00)
Process family	engagement	No detected association (Holm p = 0.874)	No detected association (Holm p = 0.484)	No detected association (Holm p = 1.00)
Vendor type	engagement	No detected association (Holm p = 1.00)	No detected association (Holm p = 1.00)	No detected association (Holm p = 1.00)
Delivery location	engagement	No detected association (Holm p = 1.00)	No detected association (Holm p = 1.00)	No detected association (Holm p = 1.00)
Class of work	engagement	No detected association (Holm p = 1.00)	No detected association (Holm p = 1.00)	No detected association (Holm p = 1.00)
Respondent role	respondent	Associated (Holm p < 0.001)	Associated (Holm p < 0.001)	Associated (Holm p < 0.001)
Engagement status and lifecycle stage	engagement	Associated (Holm p < 0.001)	Associated (Holm p < 0.001)	Associated (Holm p < 0.001)
Expected cancellation	engagement	Associated (Holm p < 0.001)	Associated (Holm p < 0.001)	Associated (Holm p < 0.001)
Intent to leave	respondent	Associated (Holm p < 0.001)	Associated (Holm p < 0.001)	Associated (Holm p < 0.001)
Delivery team size	engagement	Within ±0.30 (r = −0.11)	Within ±0.30 (r = 0.09)	Within ±0.30 (r = 0.06)
Engagement age	engagement	Within ±0.30 (r = 0.07)	Within ±0.30 (r = −0.06)	Within ±0.30 (r = −0.14)

Table 3.2. Segments and study measures against the three measures of reported governance weakness (continued)

Segment or study measure	Unit	Score	Distressed governance ratings	Written instances
Cancellation	engagement	Associated (r = −0.72)	Associated (r = 0.76)	Associated (r = 0.64)
Customer satisfaction, active engagements	engagement	Associated (r = 0.80)	Associated (r = −0.74)	Associated (r = −0.55)
Engagement indicator mean without the 47	engagement	Associated (r = 0.94)	Associated (r = −0.91)	Associated (r = −0.77)

Note: Holm adjustment controls multiple comparisons within the stated categorical family, covering its three measures and five category means. Welch tests compare means without assuming equal variances. A p value describes evidence against the test's null under its assumptions, not practical importance. Correlations use engagement-level paired bases, normally 150 engagements and 94 for customer satisfaction. CI denotes a confidence interval. The recorded equivalence margin is r = ±0.30. Equivalence tests whether an association lies within that range. The approximate detectable correlation of 0.23 at n = 150 is a different quantity. A nonsignificant conventional test does not establish equivalence. Omega squared, ω^2, estimates the share of variation associated with a group distinction. AUC is the area under a receiver operating characteristic curve, describing discrimination between canceled and active engagements in this sample. Partial r is the correlation after accounting for a specified comparison variable. Here that variable is the mean of the 53 indicators outside the score. Nominal p values are uncorrected for multiple testing. Cluster-robust inference allows standard errors to reflect dependence among respondents on the same engagement.

1 The Governance, Oversight, and Control Score

The score and low-rating counts identify different levels of weakness. Mean scores were 3 or below for 283 of 850 respondents and 50 of 150 engagements, while at least one low governance rating was recorded by 743 respondents. The composite includes controls from all five failure categories, with unequal category representation. That composition must be considered when an individual item is compared with the wider control environment.

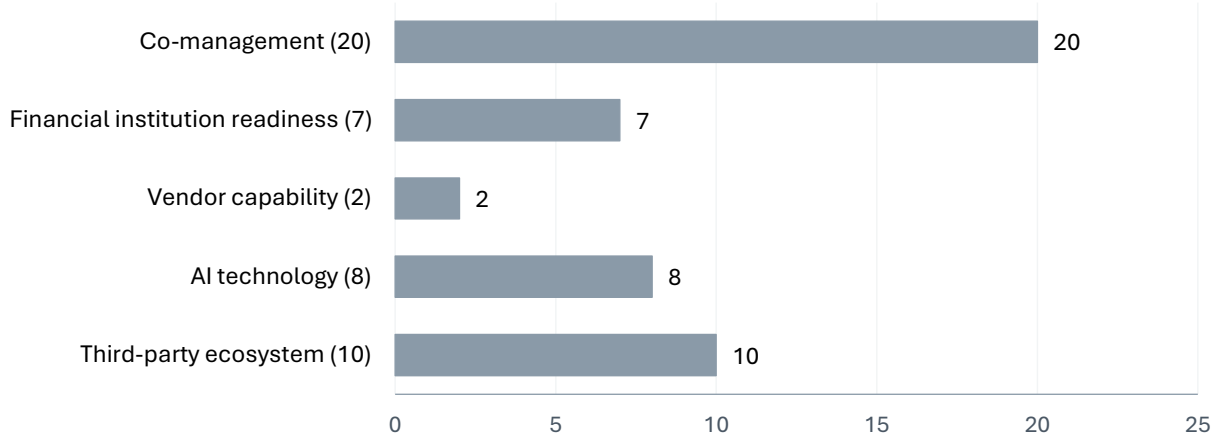
1.1 Composition of the Governance, Oversight, and Control Score

Definition · The governance, oversight, and control score · Analyzed only in Part 3

The governance, oversight, and control score places joint control between the financial institution and vendor at the center of its measurement. Co-management supplies 20 of the 47 indicators, compared with seven for financial institution readiness, two for vendor capability, eight for AI technology, and ten for the third-party ecosystem. This composition makes cross-category comparison necessary when asking whether an individual control is associated with governance elsewhere. It also means that the composite is not an equal-weight comparison of the five categories.

Figure 3.1. The 47 governance indicators, by failure category

The governance, oversight, and control score places joint control between the financial institution and vendor at the center of its measurement.



The 47 indicators divide by failure category into co-management with 20, the third-party ecosystem with 10, AI technology with 8, financial institution readiness with 7, and vendor

capability with 2. Co-management carries 42.55% of the measure. The 19 governance indicators listed in Appendix C are a subset of the 47.

Each rating entered the measure on its favorable end. The ratings were averaged for each respondent, the respondent values were averaged within each engagement, and the engagement values were averaged across the 150 engagements. Where the indicator under test was one of the 47, it was removed from the score before the comparison, so no indicator was tested against a measure that contained it.

Excluding an item from the score removes its direct contribution but leaves other related items, including 19 co-management items when a co-management indicator is tested. The cross-category comparison instead removes that entire category. Adjustment then accounts for one variable, the mean of the 53 indicators outside the score. These specifications ask different conditional questions. Neither isolates an independently established causal effect.

Table 3.3. The 47 governance indicators, by failure category

Failure category	Indicator	Part 2 section
Co-management	Co-Management of the Engagement	1.1
Co-management	Decision Rights Over Agent Action Without Human Approval	1.2
Co-management	Ownership of Each Workflow Step	1.3
Co-management	Accountability for Problems Arising From an Agent's Action	1.4
Co-management	A Single Shared View of Agent Risk	2.1
Co-management	Co-Management of Issues Arising From Agent Work	2.2
Co-management	Resolution Speed Against the Spread of Agent-Caused Problems	2.3
Co-management	Speed of Stopping a Misbehaving Agent in Live Operations	2.4
Co-management	Joint Change Control Against the Speed of Agent Behavior Change	3.1

Table 3.3. The 47 governance indicators, by failure category (continued)

Failure category	Indicator	Part 2 section
Co-management	Visibility of Agent Updates to the Other Firm Before They Take Effect	3.2
Co-management	Joint Governance of Agent Behavior Change Between Formal Releases	3.3
Co-management	Placement of Human Review Points	4.1
Co-management	Human Review Capacity Against Agent Output Volume	4.2
Co-management	Genuine Scrutiny at Human Review Points	4.3
Co-management	Reliability of the Joint Quality Assurance Loop	4.4
Co-management	Quality Measures That Capture Agent Errors	4.5
Co-management	Each Firm's Visibility of the Other Firm's Agent Actions	5.1
Co-management	Joint Decision Speed Against Agent Speed	5.2
Co-management	Fit of the Joint Governance Model to Work Performed by Agents	5.3
Co-management	Readiness of the Co-Management Model for Expected Agent Autonomy	5.4
Financial institution readiness	Maturity of the Financial Institution's Data Governance	1.4

Table 3.3. The 47 governance indicators, by failure category (continued)

Failure category	Indicator	Part 2 section
Financial institution readiness	Effectiveness of the Financial Institution's Management of Its Side	5.1
Financial institution readiness	Speed of the Financial Institution's Response to Issues Raised	5.2
Financial institution readiness	Assignment of Ownership and Accountability for AI	6.1
Financial institution readiness	Capacity to Govern and Oversee Agent Risk	6.2
Financial institution readiness	Change Control for the Changes the Engagement Requires	6.3
Financial institution readiness	Ability to Provision Access and Identities for Agents	7.1
Vendor capability	Maturity of the Playbook From Concept to Governed Production	6.1
Vendor capability	Readiness to Take Accountability for a Costly Agent Error	7.1
AI technology	Resistance to Hostile Embedded Instructions	5.3
AI technology	Safe Handling of Sensitive Data	5.2
AI technology	Consistency of Intended and Authorized Agent Action	6.1
AI technology	Reliability of Built-In Guardrails in Production	6.2
AI technology	Proof of Correct Behavior Before Work Goes Live	7.1
AI technology	Accuracy of the Tools' Own Success and Failure Reports	8.1
AI technology	Explanation and Evidence of Outputs at Audit Depth	8.2

Table 3.3. The 47 governance indicators, by failure category (continued)

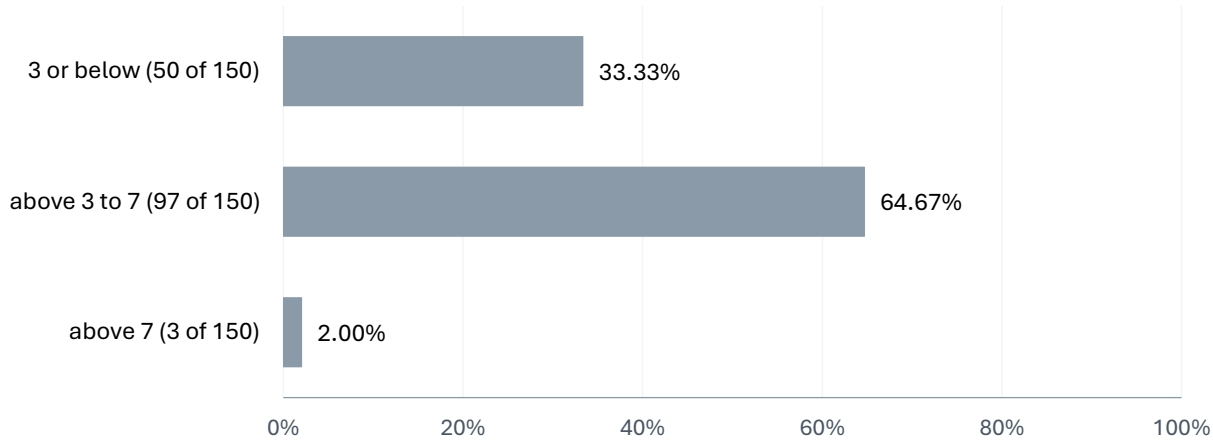
Failure category	Indicator	Part 2 section
AI technology	Quality of Logs, Traces, and Usage Telemetry	8.3
Third-party ecosystem	Quality of Advance Notice From Third-Party Providers	1.1
Third-party ecosystem	Delivery of Change Notices to the Operators of Affected Agents	1.2
Third-party ecosystem	Protection of Agents From Silent Failure	2.1
Third-party ecosystem	Monitoring for Unannounced Third-Party Change	2.2
Third-party ecosystem	Contractual Coverage of Disruptive Provider Change	4.1
Third-party ecosystem	Security Posture of Third-Party Tools and Sub-Processors	6.1
Third-party ecosystem	Management of Dependence on a Small Number of Model and Cloud Providers	6.2
Third-party ecosystem	Replacement of Informal Human Fixes With Explicit Checks	7.4
Third-party ecosystem	Mapping of Third- and Fourth-Party Dependencies	8.1
Third-party ecosystem	Management of Dependence on Gated External Rails	8.2

1.2 The Score Across Respondents and Engagements

Definition · The governance, oversight, and control score across respondents and engagements · Analyzed only in Part 3

Low ratings were more widespread than low composite means. A respondent could report a serious weakness in one control while giving other controls higher ratings, so the mean alone does not describe the reach of reported deficiencies. The distinction matters for oversight: isolated weaknesses and a broadly low assessment of controls are different findings. Both remain visible through separate reporting of low individual ratings, respondent means, and engagement means.

Figure 3.2. Share of engagements, by numeric score band



Of the 37,472 ratings given to the 47 indicators, 16,074 (42.90%) were distressed, and they came from 743 respondents on 145 engagements. The score stood at 3.99 across respondents (n = 850) and at 4.01 across engagements (n = 150). Its distressed range held 283 respondents (33.29%) and 50 engagements (33.33%).

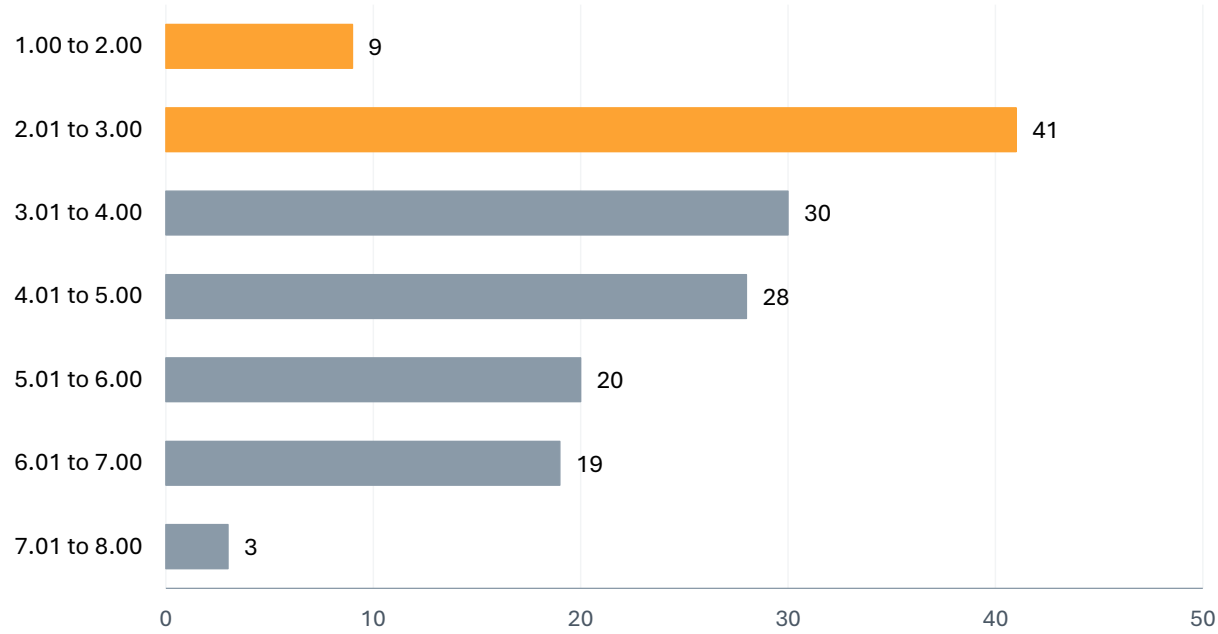
Favorable indicator ratings usually coincided with composite scores above 3. Among respondents rating a governance indicator at 8–10, none had a score of 3 or below after that item was omitted, with one exception among the 60 rating tool-report accuracy at 8–10. Composite scores were at 3 or below for 35.83% of individual contributors, 28.00% of engagement managers, and 6.45% of portfolio directors. The mean score was 2.58 on canceled engagements, compared with 4.69–4.94 across the active stages.

Written instances reached further than either measure. They numbered 3,330 of the 6,862 written answers that addressed the respondent's own engagement (48.53%), from 838 of the 850 respondents on all 150 engagements. The five engagements without a distressed governance rating carried between 5 and 15 written instances each. On one of them, an individual contributor reported correcting almost every batch without raising a ticket, the correction accepted by the team as part of the job. On another, an individual contributor reported that the ownership of a case taken over midway by an agent had never been written down.

> *"This happens on almost every batch, because of that i do not raise ticket for each one, only when it affects release, for that reason most of correcting stays with me and team just accepts it as part of job"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

> *"Happy enough with this one in truth. But one thing is missing here... when merchant raises chargeback dispute and agent picks it up halfway, nobody has written down who owns case after that point. My team decides between ourselves, usually works out. And nobody has ever asked us how we decide it"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Payments, cards and disputes

Figure 3.3. Engagements by the score, 150 engagements



Note: Amber identifies the 50 engagements with a mean governance, oversight, and control score of 3 or below. The threshold is descriptive and is not a validated diagnosis of overall governance failure.

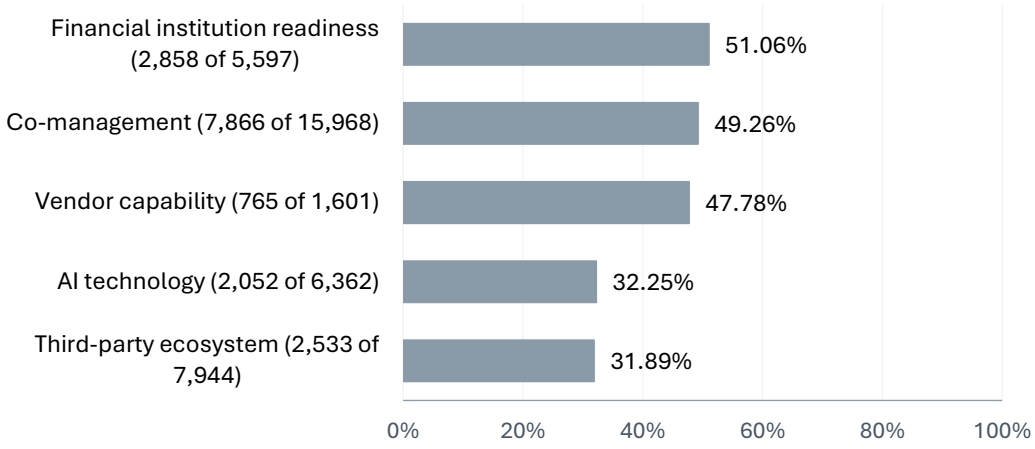
> *Data governance gaps in the source feeds were logged by the vendor as agent exceptions rather than data issues, which made the bot look like the problem when the feed was the actual cause."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Trade, settlement and reconciliation

2 Governance Indicators Across the Five Failure Categories

The five categories showed related governance concerns at different levels of low-rating prevalence. Readiness, co-management, and vendor capability had larger low-rating shares than the technology and third-party categories. This comparison concerns pooled ratings rather than unique respondents. The cross-category analysis asks a further question: whether an individual control is associated with the wider score after accounting for the other-indicator mean. Its interpretation depends on category composition and test-family choice.

Figure 3.4. Share of governance ratings in the distressed range, by failure category

The five categories showed related governance concerns at different levels of low-rating prevalence.



Low indicator ratings and a low composite mean are different quantities. Among respondents rating a technology control at 1–3, 74.90%–78.54% had a score of 3 or below after the tested indicator was omitted. The corresponding range was 73.41%–80.41% for third-party controls, 61.19%–65.21% for co-management, and 54.26%–57.31% for readiness. These conditional shares are not the prevalence of low indicator ratings. Table 3.4 reports the separate cross-category associations under the specified model.

Table 3.4. The five failure categories against governance in the other four categories

Failure category	Indicators	A score of 3 or below among respondents rating an indicator 1 to 3	r with the score without the indicator	r with the other four categories	Adjusted partial r	Indicators at nominal p < 0.05
Financial institution readiness	7	54.26% to 57.31%	0.75 to 0.80	0.72 to 0.76	−0.18 to 0.22	2 of 7
Co-management	20	61.19% to 65.21%	0.90 to 0.93	0.75 to 0.80	0.21 to 0.31	20 of 20
Vendor capability	2	59.18% to 64.34%	0.78 to 0.79	0.78 to 0.78	−0.46 to −0.38	2 of 2
AI technology	8	74.90% to 78.54%	0.82 to 0.86	0.77 to 0.81	−0.09 to 0.08	0 of 8
Third-party ecosystem	10	73.41% to 80.41%	0.83 to 0.85	0.75 to 0.78	−0.14 to −0.04	0 of 10

Note: Correlations use 150 engagements. For an item-to-score comparison, the score excludes that item. For a cross-category comparison, it excludes the item's entire failure category. Partial r is the correlation after accounting for one comparison measure, the mean of the 53 indicators outside the score. It does not estimate 53 separate adjustment coefficients. Cross-category p values are nominal unless an adjustment is expressly identified. Statistical significance at nominal $p < 0.05$ should not be confused with a family-adjusted conclusion. Category composition differs in both the score and the comparison mean.

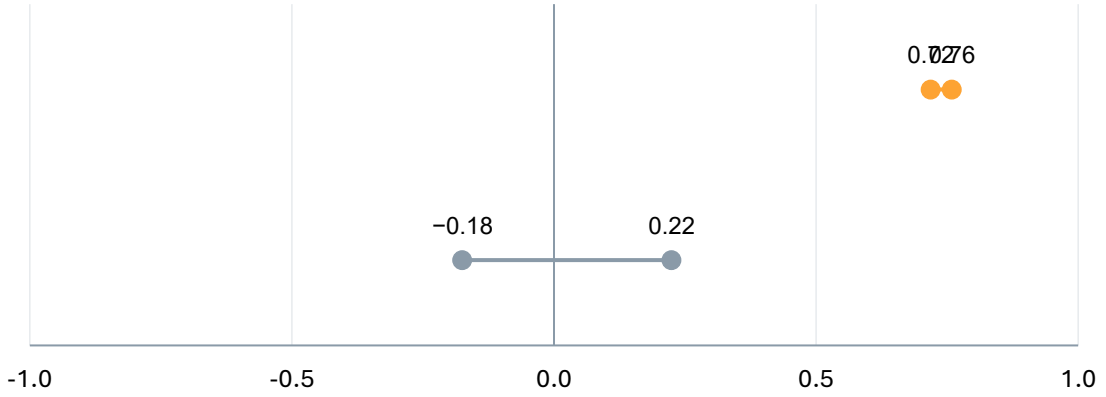
2.1 Financial Institution Readiness

Failure category · 7 governance indicators · Full analysis of each indicator in Part 2, Financial Institution Readiness

Financial institution readiness had the largest share of low governance-related ratings. This places the institution's own data, access, management, and control responsibilities within the governance problem rather than treating governance as a vendor-only obligation. The adjusted comparisons provide more qualified evidence about relationships with the other categories. They cannot establish that the institution's individual controls matter only through its overall readiness.

Figure 3.5. Financial Institution Readiness: ranges of unadjusted and adjusted cross-category associations.

Financial institution readiness associations weakened after adjustment



The endpoints show the minimum and maximum coefficients across indicators. These ranges are not confidence intervals.

Comparison	Estimate and analysis base
Unadjusted minimum r	r = 0.72 150 engagements
Unadjusted maximum r	r = 0.76 150 engagements
Adjusted minimum partial r	partial r = −0.18 150 engagements
Adjusted maximum partial r	partial r = 0.22 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

The score excludes the tested category. Adjustment uses the mean of the other 53 indicators. Ranges describe estimates across items, not confidence intervals. Category-mean bands are 3 or below: 60 of 150 (40.00%). Above 3 to 7: 86 of 150 (57.33%). Above 7: 4 of 150 (2.67%).

The seven readiness indicators were distressed in 2,858 of 5,597 ratings (51.06%), the highest share of the five categories, and their means ran from 3.47 to 3.61. The category mean was distressed on 60 of the 150 engagements (40.00%).

Each readiness indicator was associated with governance in the other four categories at r = 0.72 to 0.76 (n = 150 engagements). After adjustment for the mean of the other 53 indicators, the partial r ran from −0.18 to 0.22, and two of the seven associations were significant. The governance of the financial institution's data carried the one positive tie (partial r = 0.22). The speed of the financial institution's response to issues raised carried a small inverse result (partial r = −0.18) that reversed to 0.13 once the institution's other readiness indicators were omitted from the comparison mean.

The association of each indicator with the score computed without it, measured across respondents, held within the canceled engagements at r = 0.54 to 0.62, the lowest of the five categories. It held within each active stage, at 0.58 to 0.69 in Build-Out, 0.71 to 0.79 in Transition/Cut-Over, and 0.64 to 0.76 in Production. Within role, it ran from 0.74 to 0.78 among individual contributors and from 0.51 to 0.75 among the 31 portfolio directors.

Among respondents who rated a readiness indicator 1 to 3, 54.26% to 57.31% had a score of 3 or below, the lowest range of the five categories. Across the seven indicators, ratings of 8 to 10 numbered 126, and none came from a respondent with a score of 3 or below. The 1,479 written answers on the readiness indicators held 1,323 negative answers, and 381 of the negative answers described one of the 26 forms of a control bypassed or not working.

Part 2 (Financial Institution Readiness) analyzes each indicator in full, and Financial Institution Readiness and the Governance of Agentic Work reads each against reported governance weakness.

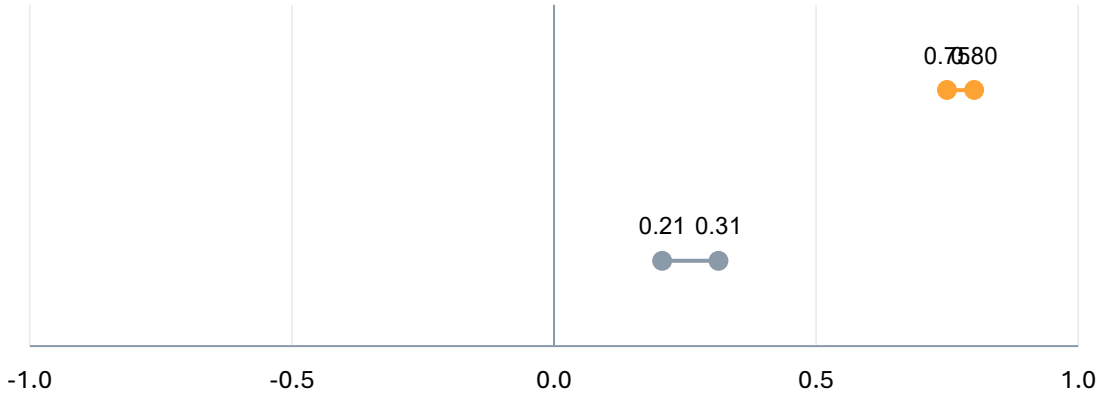
2.2 Co-Management

Failure category · 20 governance indicators · Full analysis of each indicator in Part 2, Co-Management

Co-management showed consistently positive associations with governance in the other four categories, including after the specified composite adjustment. The substantive concern is coordination across firms: decision rights, change control, review, and incident management must operate across the same work. The consistency of the 20 estimates supports examining that shared operating arrangement. Its apparent statistical distinctiveness remains conditional on score composition, comparison-measure composition, and the stated family of tests.

Figure 3.6. Co-Management: ranges of unadjusted and adjusted cross-category associations.

Co-management showed consistently positive associations with governance in the other four categories, including after the specified composite adjustment.



The endpoints show the minimum and maximum coefficients across indicators. These ranges are not confidence intervals.

Comparison	Estimate and analysis base
Unadjusted minimum r	r = 0.75 150 engagements
Unadjusted maximum r	r = 0.80 150 engagements
Adjusted minimum partial r	partial r = 0.21 150 engagements
Adjusted maximum partial r	partial r = 0.31 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

The score excludes the tested category. Adjustment uses the mean of the other 53 indicators. Ranges describe estimates across items, not confidence intervals. Category-mean bands are 3 or below: 63 of 150 (42.00%). Above 3 to 7: 83 of 150 (55.33%). Above 7: 4 of 150 (2.67%).

Each of its 20 indicators was associated with governance in the other four categories at r = 0.75 to 0.80 (n = 150 engagements). After that adjustment, the partial r ran from 0.21 to 0.31, and all 20 associations were significant. No indicator outside co-management carried a significant positive adjusted association except the governance of the financial institution's data.

The 20 indicators were distressed in 7,866 of 15,968 ratings (49.26%), the second-highest share of the five categories, and their means ran from 3.60 to 3.73. The category mean was distressed on 63 of the 150 engagements (42.00%), more than in any other category.

The association of each indicator with the score computed without it held within the canceled engagements at r = 0.74 to 0.82, the highest of the five categories. Within the active stages it ran from 0.75 to 0.83 in Build-Out, 0.80 to 0.86 in Transition/Cut-Over, and 0.82 to 0.89 in Production. Within role, it ran from 0.86 to 0.88 among individual contributors and from 0.76 to 0.90 among portfolio directors.

Among respondents who rated a co-management indicator 1 to 3, 61.19% to 65.21% had a score of 3 or below. Across the 20 indicators, ratings of 8 to 10 numbered 605, and none came from a respondent with a score of 3 or below. The 1,146 written answers on the co-management indicators held 903 negative answers, and 199 of the negative answers described one of the 26 forms of a control bypassed or not working.

Part 2 (Co-Management) analyzes each indicator in full, and Co-Management and Governance Across the Other Four Categories reads each against reported governance weakness.

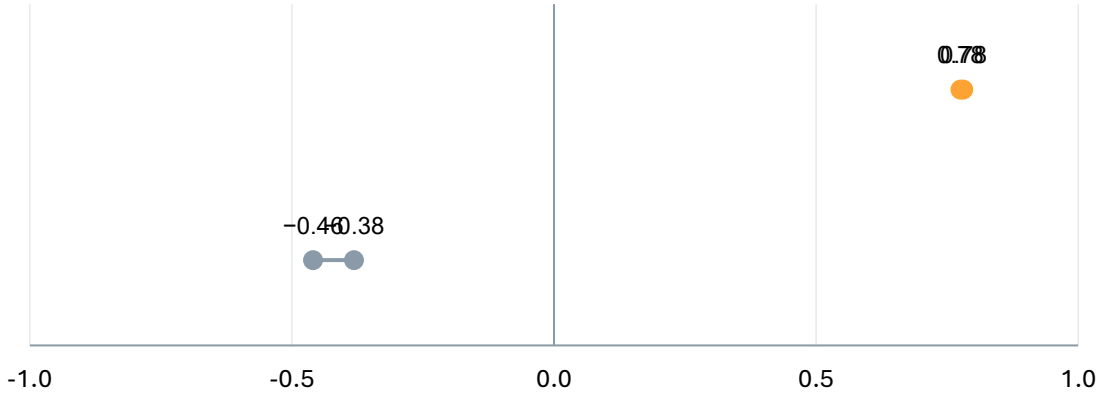
2.3 Vendor Capability

Failure category · 2 governance indicators · Full analysis of each indicator in Part 2, Vendor Capability

The vendor's playbook and readiness to accept accountability were associated with the wider control environment, but their adjusted coefficients changed sharply with the comparison measure. For governance, these indicators concern whether the vendor has a defined route into controlled production and accepts responsibility when agent work causes loss. The results support examining those responsibilities. They do not show that weak vendor controls improve governance or that their relationship is exhausted by general capability.

Figure 3.7. Vendor Capability: ranges of unadjusted and adjusted cross-category associations.

Vendor control associations changed substantially after adjustment



The endpoints show the minimum and maximum coefficients across indicators. These ranges are not confidence intervals.

Comparison	Estimate and analysis base
Unadjusted minimum r	r = 0.78 150 engagements
Unadjusted maximum r	r = 0.78 150 engagements
Adjusted minimum partial r	partial r = −0.46 150 engagements
Adjusted maximum partial r	partial r = −0.38 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

The score excludes the tested category. Adjustment uses the mean of the other 53 indicators. Ranges describe estimates across items, not confidence intervals. Category-mean bands are 3 or below: 58 of 150 (38.67%). Above 3 to 7: 88 of 150 (58.67%). Above 7: 4 of 150 (2.67%).

The two indicators were distressed in 765 of 1,601 ratings (47.78%), and their means were 3.65 and 3.67. The category mean was distressed on 58 of the 150 engagements (38.67%).

The two vendor indicators correlated with governance in the other four categories at approximately r = 0.78. Their adjusted partial correlations were −0.38 for the playbook and −0.46 for accountability, both nominally significant. The comparison mean contains the vendor's 18 other capability indicators. Removing those items changed the estimates to 0.04 and −0.03. This sensitivity demonstrates dependence on the adjustment measure and limits an interpretation of the negative coefficients as distinct effects of these controls.

The association of each indicator with the score computed without it held within the canceled engagements at r = 0.66 to 0.69. Within the active stages it ran from 0.70 to 0.71 in Build-Out, 0.65 to 0.66 in Transition/Cut-Over, and 0.77 to 0.81 in Production. Within role, it ran from 0.78 to 0.79 among individual contributors and stood at 0.68 for each indicator among portfolio directors.

Among respondents who rated a vendor capability indicator 1 to 3, 59.18% and 64.34% had a score of 3 or below. Ratings of 8 to 10 numbered 44 across the two indicators, and none came from a respondent with a score of 3 or below. The 238 written answers on the two indicators held 215 negative answers, and 25 of the negative answers described one of the 26 forms of a control bypassed or not working.

Part 2 (Vendor Capability) analyzes each indicator in full, and Vendor Preparation and Accountability in the Control Environment reads both against reported governance weakness.

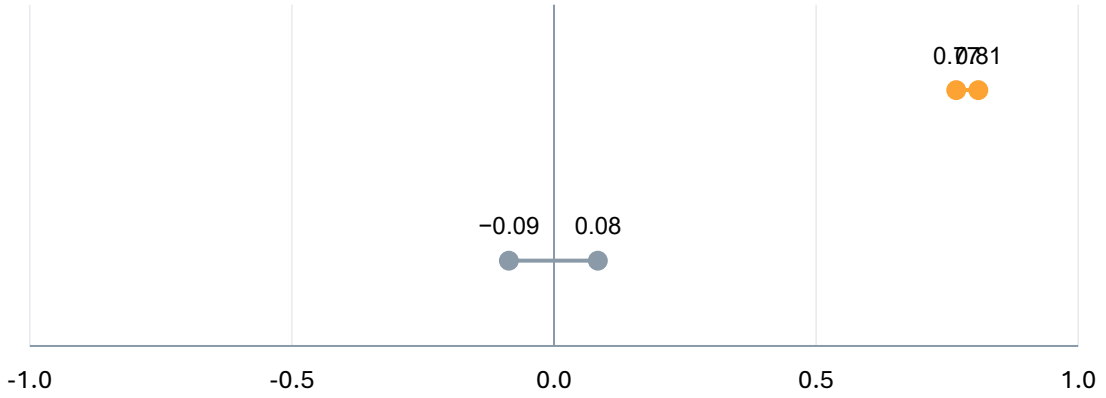
2.4 AI Technology

Failure category · 8 governance indicators · Full analysis of each indicator in Part 2, AI Technology

Low technology-control ratings were associated with low composite scores, connecting the operation of technical safeguards to the broader control environment. The eight indicators concern whether agent actions remain authorized, observable, testable, and subject to effective safeguards. The main adjusted model did not detect positive cross-category associations for these controls. That conditional result limits a claim of statistical independence, while leaving their substantive control functions and strong observed associations intact.

Figure 3.8. AI Technology: ranges of unadjusted and adjusted cross-category associations.

Technology-control associations weakened after adjustment



The endpoints show the minimum and maximum coefficients across indicators. These ranges are not confidence intervals.

Comparison	Estimate and analysis base
Unadjusted minimum r	r = 0.77 150 engagements
Unadjusted maximum r	r = 0.81 150 engagements
Adjusted minimum partial r	partial r = −0.09 150 engagements
Adjusted maximum partial r	partial r = 0.08 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

The score excludes the tested category. Adjustment uses the mean of the other 53 indicators. Ranges describe estimates across items, not confidence intervals. Category-mean bands are 3 or below: 30 of 150 (20.00%). Above 3 to 7: 104 of 150 (69.33%). Above 7: 16 of 150 (10.67%).

The eight technology indicators were distressed in 2,052 of 6,362 ratings (32.25%), and their means ran from 4.54 to 4.66. The category mean was distressed on 30 of the 150 engagements (20.00%).

Among respondents who rated a technology control 1 to 3, 74.90% to 78.54% had a score of 3 or below. Across the eight indicators, ratings of 8 to 10 numbered 551, and one came from a respondent with a score of 3 or below.

The technology indicators correlated with the other four categories at r = 0.77–0.81. In the main adjusted model, partial r ranged from −0.09 to 0.08 and none met nominal $p < 0.05$. This is a conditional result. For authorized action, removing the other 12 AI technology items from the comparison mean changed partial r from −0.03 ($p = 0.745$) to 0.29 ($p < 0.001$). The sensitivity qualifies any conclusion that technology controls have no relationship of their own.

Associations between each technology indicator and the score excluding it ranged from 0.65 to 0.70 in Build-Out, 0.71 to 0.77 in Transition/Cut-Over, and 0.78 to 0.82 in Production. Within canceled engagements they ranged from 0.66 to 0.72. The stage ordering is descriptive and is not a test that associations strengthened over time. Within role, correlations ranged from 0.79 to 0.82 among individual contributors and from 0.61 to 0.87 among portfolio directors.

The 493 written answers on the technology indicators held 402 negative answers, and 42 of the negative answers described one of the 26 forms of a control bypassed or not working.

Part 2 (AI Technology) analyzes each indicator in full, and Technical Controls and Governance Across the Engagement reads each against reported governance weakness.

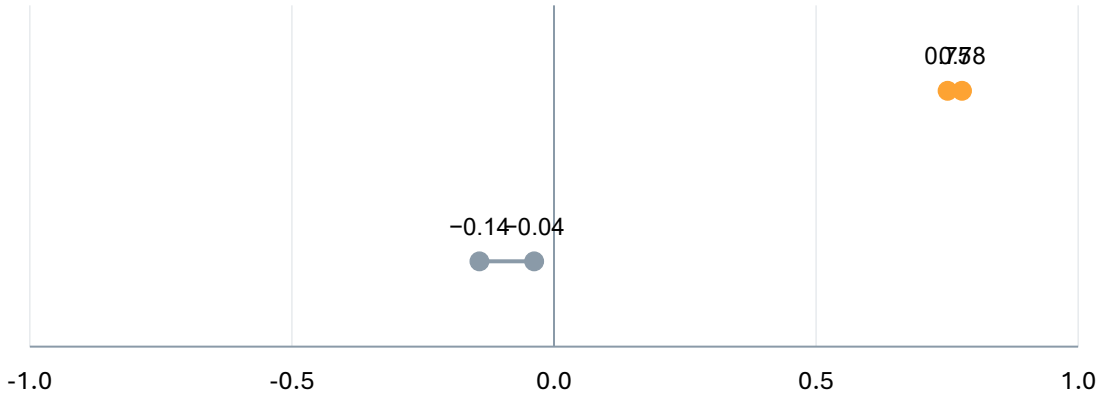
2.5 Third-Party Ecosystem

Failure category · 10 governance indicators · Full analysis of each indicator in Part 2, Third-Party Ecosystem

Third-party control ratings connected external dependencies with the broader governance environment. Advance notice, monitoring, contractual coverage, and dependency management concern work that the financial institution and vendor cannot govern through internal procedures alone. Low ratings in this category commonly accompanied low composite scores. The adjusted results did not establish a separate positive relationship under the specified model, and they do not demonstrate that external controls have no distinct operational importance.

Figure 3.9. Third-Party Ecosystem: ranges of unadjusted and adjusted cross-category associations.

Third-party control ratings connected external dependencies with the broader governance environment.



The endpoints show the minimum and maximum coefficients across indicators. These ranges are not confidence intervals.

Comparison	Estimate and analysis base
Unadjusted minimum r	r = 0.75 150 engagements
Unadjusted maximum r	r = 0.78 150 engagements
Adjusted minimum partial r	partial r = −0.14 150 engagements
Adjusted maximum partial r	partial r = −0.04 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

The score excludes the tested category. Adjustment uses the mean of the other 53 indicators. Ranges describe estimates across items, not confidence intervals. Category-mean bands are 3 or below: 34 of 150 (22.67%). Above 3 to 7: 100 of 150 (66.67%). Above 7: 16 of 150 (10.67%).

The ten third-party indicators were distressed in 2,533 of 7,944 ratings (31.89%), the lowest share of the five categories, and their means ran from 4.58 to 4.66. The category mean was distressed on 34 of the 150 engagements (22.67%).

Among respondents who rated a third-party control 1 to 3, 73.41% to 80.41% had a score of 3 or below, and the upper end of the range was the highest share of the 47 governance indicators. Across the ten indicators, ratings of 8 to 10 numbered 764, and none came from a respondent with a score of 3 or below.

Each third-party indicator was associated with governance in the other four categories at r = 0.75 to 0.78 (n = 150 engagements). After adjustment for the mean of the other 53 indicators, the partial r ran from −0.14 to −0.04, and none of the ten was significant.

Within canceled engagements, third-party indicator associations with the score excluding the item ranged from 0.65 to 0.73. Across active stages, the ranges were 0.70–0.79 in Build-Out, 0.69–0.77 in Transition/Cut-Over, and 0.72–0.81 in Production. These comparisons do not establish a change over time. Within role, correlations ranged from 0.79 to 0.82 among individual contributors and from 0.70 to 0.84 among portfolio directors.

The 522 written answers on the third-party indicators held 372 negative answers, and 233 of the negative answers described one of the 26 forms of a control bypassed or not working, the highest proportion of the five categories.

Part 2 (Third-Party Ecosystem) analyzes each indicator in full, and Third-Party Dependencies and the Limits of Internal Control reads each against reported governance weakness.

3 Governance Across the Segments of the Study

Structural classifications did not clearly distinguish governance depth after the stated adjustment, whereas respondent role did. Reported weaknesses extended across financial institution types, outsourcing models, process families, vendor types, locations, and classes of work. This pattern gives limited support to a simple structural explanation of the score variation. It also directs attention to the operation of controls within engagements, without establishing equality between the structural groups.

Table 3.5. The score, distressed governance ratings, and written instances of governance weakness, by segment

Segment	Engagements	Score	Distressed governance ratings per engagement	Written instances per engagement
Financial institution type				
Asset management	24	3.85	111.33	21.88
Wealth management and private banking	23	4.07	114.22	23.30
Investment banking (advisory and capital-markets origination)	19	3.34	134.42	22.58
Insurance (life, property and casualty, health)	14	3.95	96.86	22.00
Commercial and business banking	13	4.38	100.54	21.54
Payments and merchant acquiring (networks, acquiring, processing)	13	4.28	96.23	20.54

Table 3.5. The score, distressed governance ratings, and written instances of governance weakness, by segment (continued)

Segment	Engagements	Score	Distressed governance ratings per engagement	Written instances per engagement
Credit card and payments issuing	12	3.81	125.42	25.00
Retail and consumer banking	12	4.40	97.83	23.25
Markets and trading (sales and trading, global markets)	11	4.38	82.55	23.00
Private equity and private capital	9			
Outsourcing model				
CXM	42	4.28	90.71	21.07
ITO	40	4.07	108.45	24.57
BPO	37	3.64	122.22	22.70
KPO	31	3.98	109.81	20.06
Vendor type				
IT services major	59	4.09	99.03	21.36
BPO specialist	37	3.68	126.81	23.08
Consulting firm	23	4.28	101.26	21.30
Boutique or specialist provider	17	4.06	100.59	23.88
Pure-play agentic AI firm	14	4.02	107.14	22.86
Delivery location				
India	92	4.03	108.07	22.46
United States	19	3.79	113.89	23.21
All other locations	39	4.07	101.74	21.10
Class of work				
LOB Operational	76	3.93	117.26	22.86
Corporate Functional	74	4.08	96.78	21.53
Engagement status and lifecycle stage				
Build-Out	31	4.85	57.39	18.55
Transition/Cut-Over	43	4.94	62.37	19.47

Table 3.5. The score, distressed governance ratings, and written instances of governance weakness, by segment (continued)

Segment	Engagements	Score	Distressed governance ratings per engagement	Written instances per engagement
Production	20	4.69	65.95	17.25
Canceled	56	2.58	183.82	28.09
Cancellation risk				
High risk	19	3.91	97.16	21.05
Moderate risk	39	4.63	70.69	19.21
Below moderate risk	36	5.61	32.69	16.89
Canceled	56	2.58	183.82	28.09
Indicator quintile				
Quintile 1	30	2.13	201.63	29.27
Quintile 2	30	2.90	167.17	26.10
Quintile 3	30	3.79	109.80	21.47
Quintile 4	30	4.97	45.17	19.87
Quintile 5	30	6.24	12.03	14.30

Note: Levels with fewer than 10 engagements are reported as counts only. Tests between levels ran at the engagement level. For financial institution type, the score differed with Welch $p = 0.425$ (Holm $p = 1.00$), and written instances with Welch $p = 0.618$ (Holm $p = 1.00$). For outsourcing model, the score differed with Welch $p = 0.311$ (Holm $p = 1.00$), and written instances with Welch $p = 0.614$ (Holm $p = 1.00$). For vendor type, the score differed with Welch $p = 0.645$ (Holm $p = 1.00$), and written instances with Welch $p = 0.784$ (Holm $p = 1.00$). For delivery location, the score differed with Welch $p = 0.749$ (Holm $p = 1.00$), and written instances with Welch $p = 0.956$ (Holm $p = 1.00$). For class of work, the score differed with Welch $p = 0.534$ (Holm $p = 1.00$), and written instances with Welch $p = 0.542$ (Holm $p = 1.00$). For engagement status and lifecycle stage, the score differed with Welch $p < 0.001$ (Holm $p < 0.001$), and written instances with Welch $p < 0.001$ (Holm $p < 0.001$). For cancellation risk, the score differed with Welch $p < 0.001$ (Holm $p < 0.001$), and written instances with Welch $p < 0.001$ (Holm $p < 0.001$). For indicator quintile, the score differed with Welch $p < 0.001$ (Holm $p < 0.001$), and written instances with Welch $p < 0.001$ (Holm $p < 0.001$). Omega squared (ω^2) is an estimate of the proportion of variation associated with a group distinction. Cluster-robust inference allows respondent errors to be dependent within the same engagement. Nonsignificance is not equivalence.

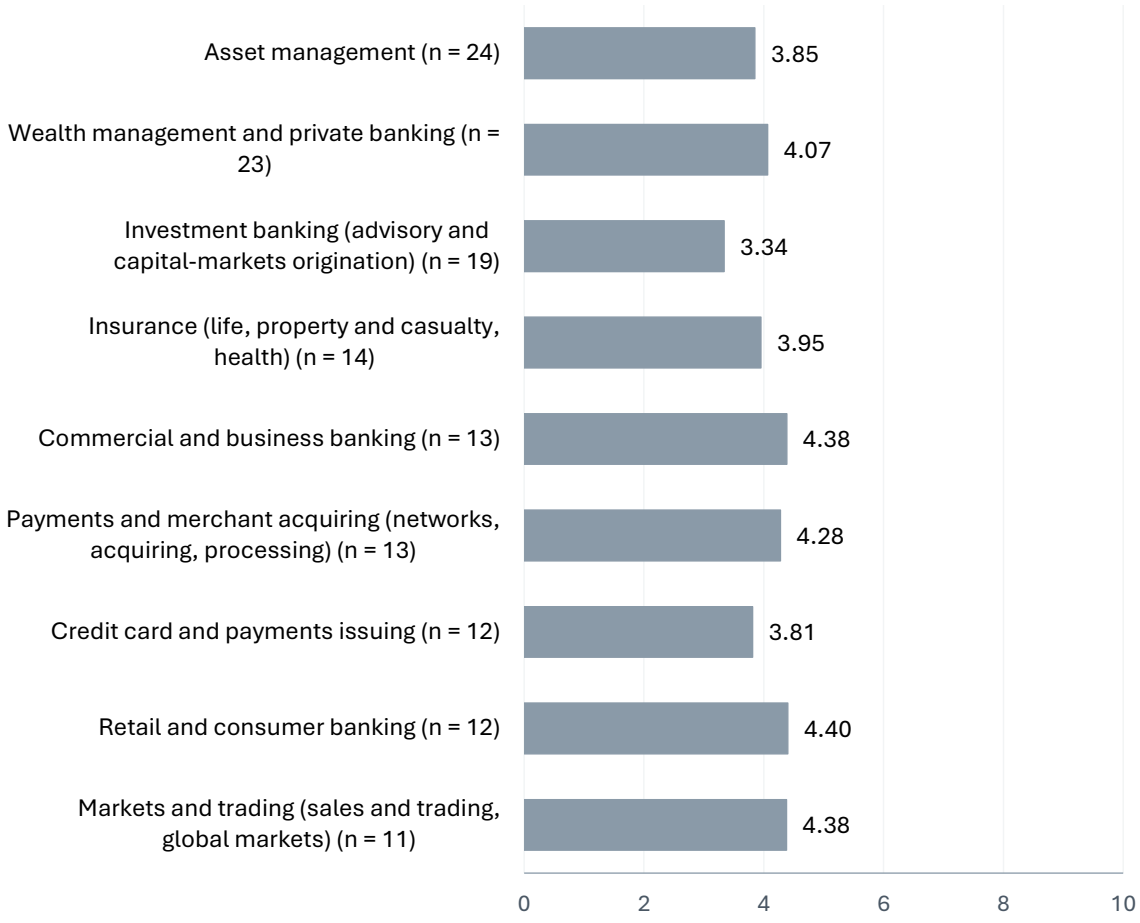
3.1 Financial Institution Type

Segment · 10 financial institution types · Full analysis in Part 1, Structural Attributes and Outcome Associations, section 1.1

Reported governance weaknesses extended across the financial institution types examined. The adjusted comparisons did not detect differences in the three measures of depth, while lower scores on canceled engagements appeared within the larger type groups. The evidence gives limited support for attributing the governance pattern to financial institution type. It does not establish that the types have identical control arrangements or identical levels of risk.

Figure 3.10. The score by institution type, engagements with 10 or more

Reported governance weaknesses extended across the financial institution types examined.



Across the nine types with 10 or more engagements, the score ranged from 3.34 in investment banking (n = 19 engagements) to 4.40 in retail and consumer banking (n = 12), with Holm p = 1.00. Distressed governance ratings per respondent ranged from 15.50 to 23.94, and written instances per respondent from 3.42 to 4.41, each with Holm p = 1.00.

The five failure categories showed no clearly detected adjusted differences. Co-management ran from 2.83 in investment banking to 4.15 in payments and merchant acquiring, and the third-party ecosystem from 3.64 in investment banking to 5.39 in retail and consumer banking. Each of the five category means carried Holm p = 1.00, including financial institution readiness at 3.14 to 4.00.

Within each type, the score was associated with cancellation, from r = −0.56 in payments and merchant acquiring (n = 13 engagements) to r = −0.85 in retail and consumer banking (n = 12). Private equity and private capital, with 9 engagements and a score of 4.08, is reported as a count only.

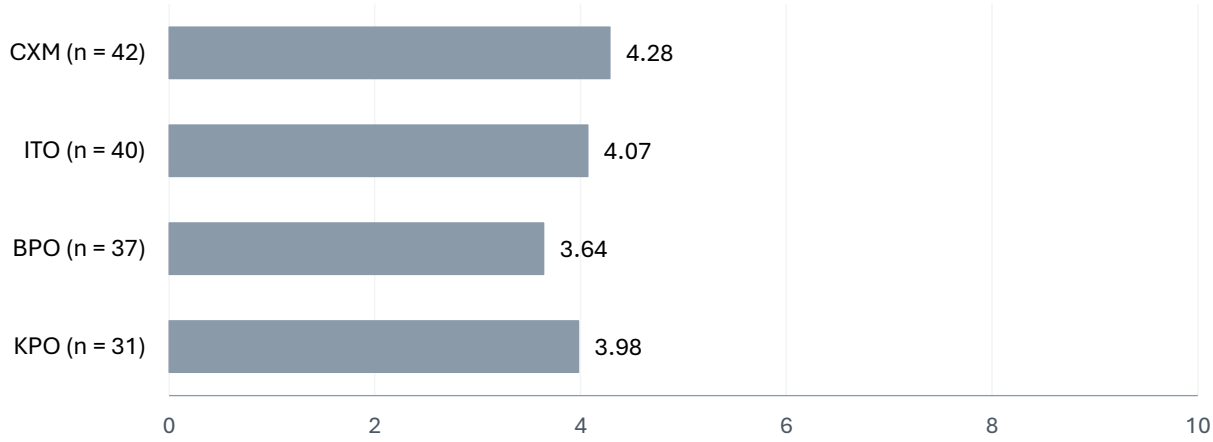
3.2 Outsourcing Model

Segment · 4 outsourcing models · Full analysis in Part 1, Structural Attributes and Outcome Associations, section 3.1

The governance pattern extended across all four outsourcing models. Each included lower-scoring canceled engagements and reported weaknesses across the five categories, without a detected adjusted difference between models. For the paper's central question, the model label alone does not account for the observed variation in control assessments. Differences in operating arrangements within a model remain possible and should not be dismissed by a nonsignificant group comparison.

Figure 3.11. The score by outsourcing model, engagements with 10 or more

The governance pattern extended across all four outsourcing models.



The score stood at 4.28 under CXM (n = 42 engagements), 4.07 under ITO (n = 40), 3.98 under KPO (n = 31), and 3.64 under BPO (n = 37), with Holm p = 1.00. Distressed governance ratings per respondent ran from 16.81 under CXM to 21.87 under BPO, and written instances per respondent from 3.74 under KPO to 4.05 under BPO, each with Holm p = 1.00.

BPO engagements stood lowest in all five failure categories without differing from the other models in any of them. Co-management ran from 3.21 under BPO to 4.02 under CXM, and AI technology from 4.20 under BPO to 4.77 under ITO, each with Holm p = 1.00.

Within each model, the score was associated with cancellation at r = −0.72 under CXM and ITO, −0.71 under KPO, and −0.78 under BPO, on 12 to 15 canceled engagements per model.

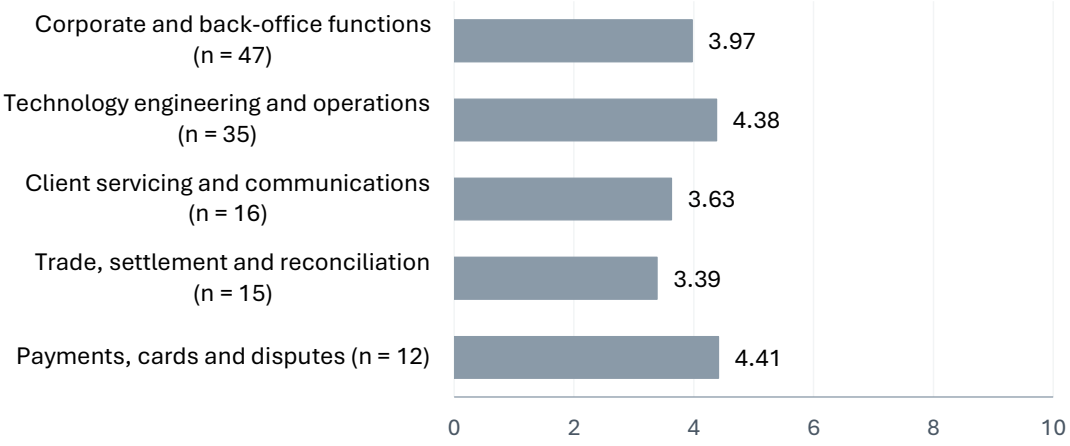
3.3 Process Family

Segment · 8 process families · Full analysis in Part 1, Structural Attributes and Outcome Associations, section 3.3

The process-family comparisons did not identify a clear adjusted difference in governance depth. This broadens the reported control concern beyond a single type of financial institution work. It also limits the explanation that the paper's overall pattern arises principally from the mix of processes sampled. The comparisons describe the families with sufficient engagement bases, with smaller groups retained descriptively.

Figure 3.12. The score by process family, engagements with 10 or more

The process-family comparisons did not identify a clear adjusted difference in governance depth.



Across the five families with 10 or more engagements, the score ranged from 3.39 in trade, settlement and reconciliation (n = 15 engagements) to 4.41 in payments, cards and disputes (n = 12), with Holm p = 0.874. Distressed governance ratings per respondent ranged from 14.37 to 24.93 (Holm p = 0.484), and written instances per respondent from 3.46 to 4.19 (Holm p = 1.00).

Co-management came closest to a difference among the eight measures. It stood at 2.76 in trade, settlement and reconciliation against 4.20 in payments, cards and disputes, a difference that reached p = 0.019 before adjustment and did not survive it (Holm p = 0.148). The share of respondents with a score of 3 or below also differed by family before adjustment (Monte Carlo p = 0.026), while the means did not.

Within each family, the score was associated with cancellation, from r = −0.63 in payments, cards and disputes to r = −0.77 in client servicing and communications. Credit, underwriting

and suitability (9 engagements, the score 4.75), regulatory, audit, tax and legal (9, 3.70), and financial crime and client onboarding (7, 3.29) fall below the threshold of 10 engagements and are reported as counts in Table 3.6.

Table 3.6. The score and written instances of governance weakness, by process family

Process family	Engagements	Respondents	Score
Corporate and back-office functions	47	248	3.97
Technology engineering and operations	35	204	4.38
Client servicing and communications	16	94	3.63
Trade, settlement and reconciliation	15	80	3.39
Payments, cards and disputes	12	73	4.41
Credit, underwriting and suitability	9	57	
Regulatory, audit, tax and legal	9	51	
Financial crime and client onboarding	7	43	

Process family	Respondents with a score of 3 or below	Written instances per engagement
Corporate and back-office functions	87 (35.08%)	21.45
Technology engineering and operations	46 (22.55%)	22.06
Client servicing and communications	41 (43.62%)	24.50
Trade, settlement and reconciliation	40 (50.00%)	20.80
Payments, cards and disputes	14 (19.18%)	21.25
Credit, underwriting and suitability		
Regulatory, audit, tax and legal		
Financial crime and client onboarding		

Note: Process families with fewer than 10 engagements are retained descriptively. The governance, oversight, and control score is the engagement-level mean. Respondent score bands refer to the respondent's mean and are descriptive. A mean or rating of 3 or below is not independently verified overall governance failure.

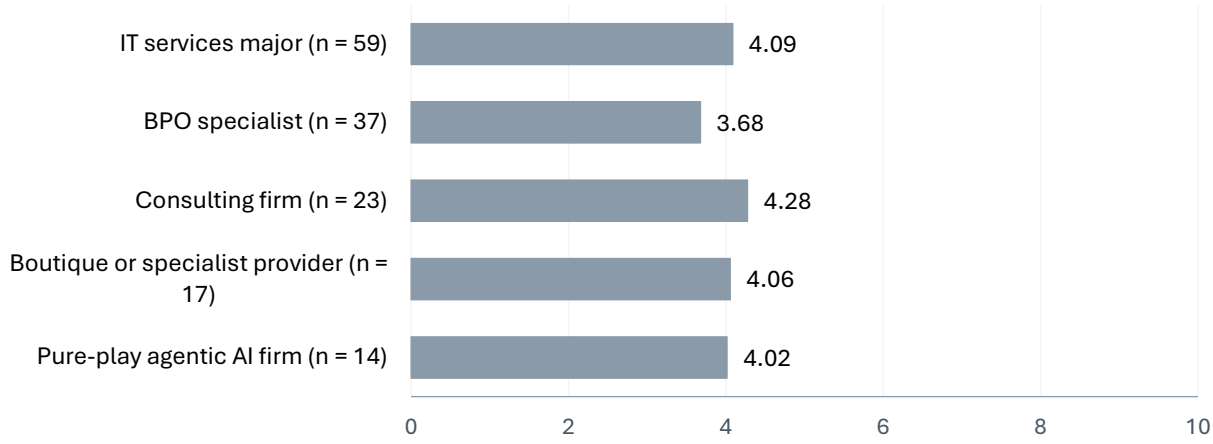
3.4 Vendor Type

Segment · 5 vendor types · Full analysis in Part 1, Structural Attributes and Outcome Associations, section 5.1

Vendor type did not clearly distinguish engagement-level governance assessments after adjustment. The nominal respondent-level difference also did not survive the reported multiplicity correction. Governance weaknesses appeared across the vendor types represented, rather than being confined to one provider model. These comparisons concern vendor classifications, not an independently measured test of each vendor's control effectiveness.

Figure 3.13. The score by vendor type (engagement), engagements with 10 or more

Vendor type did not clearly distinguish engagement-level governance assessments after adjustment.



The score ranged from 3.68 for BPO specialists (n = 37 engagements) to 4.28 for consulting firms (n = 23), with Holm p = 1.00. At the respondent level the difference reached p = 0.017 before adjustment and did not survive it (Holm p = 0.136).

Vendor capability, the category that measures the vendor itself, ran from 3.47 for BPO specialists to 4.07 for pure-play agentic AI firms, with Holm p = 1.00. BPO specialists stood lowest in all five categories, and none of the five differences survived adjustment. Distressed governance ratings per respondent ranged from 17.49 to 22.03, and written instances per respondent from 3.74 to 4.17, each with Holm p = 1.00.

Within each vendor type, the score was associated with cancellation, from r = −0.68 for BPO specialists and pure-play agentic AI firms to r = −0.83 for boutique or specialist providers. The pure-play agentic AI firms carried the fewest canceled engagements, 5 of 14.

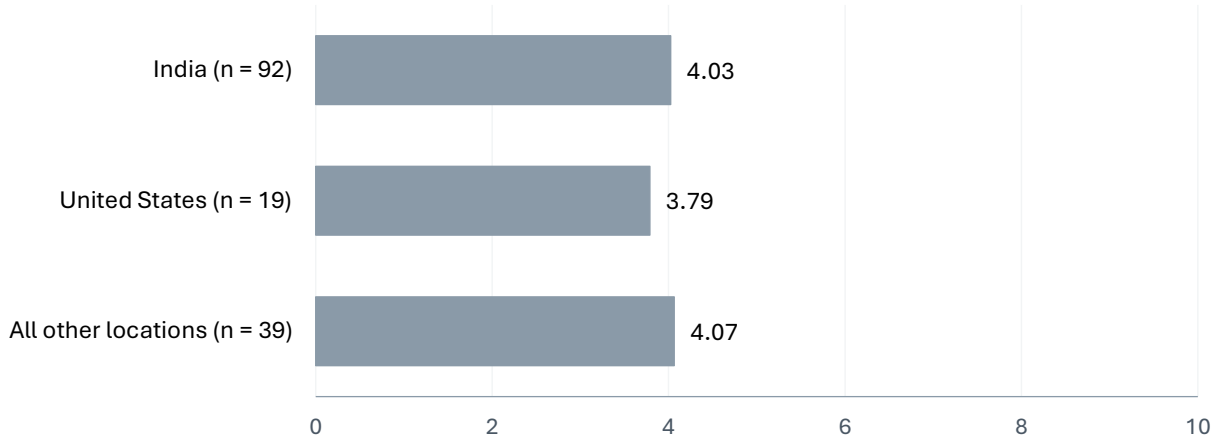
3.5 Delivery Location

Segment · 3 delivery locations · Full analysis in Part 1, Structural Attributes and Outcome Associations, section 3.4

Delivery location did not clearly distinguish the reported governance measures after adjustment. Lower scores on canceled engagements appeared within India, the United States, and the combined group of other locations. The finding directs attention to control arrangements within engagements rather than a simple geographic explanation. The combined other-location group should be understood as the reported aggregate, with no inference that every location within it is equivalent.

Figure 3.14. The score by delivery location, engagements with 10 or more

Delivery location did not clearly distinguish the reported governance measures after adjustment.



Engagements delivered from India recorded a score of 4.03 (n = 92 engagements), those delivered from the United States 3.79 (n = 19), and those delivered from all other locations 4.07 (n = 39), with Holm p = 1.00. Distressed governance ratings per respondent stood at 18.77, 19.71, and 18.67, and written instances per respondent at 3.90, 3.98, and 3.91, each with Holm p = 1.00.

The category means varied within narrow ranges. Co-management ran only from 3.58 to 3.69 across the three locations, and the widest range, in the third-party ecosystem, ran from 4.20 in the United States to 4.93 elsewhere, with Holm p = 1.00 for every category.

Within each location, the score was associated with cancellation at r = −0.70 in India, −0.78 in the United States, and −0.74 elsewhere.

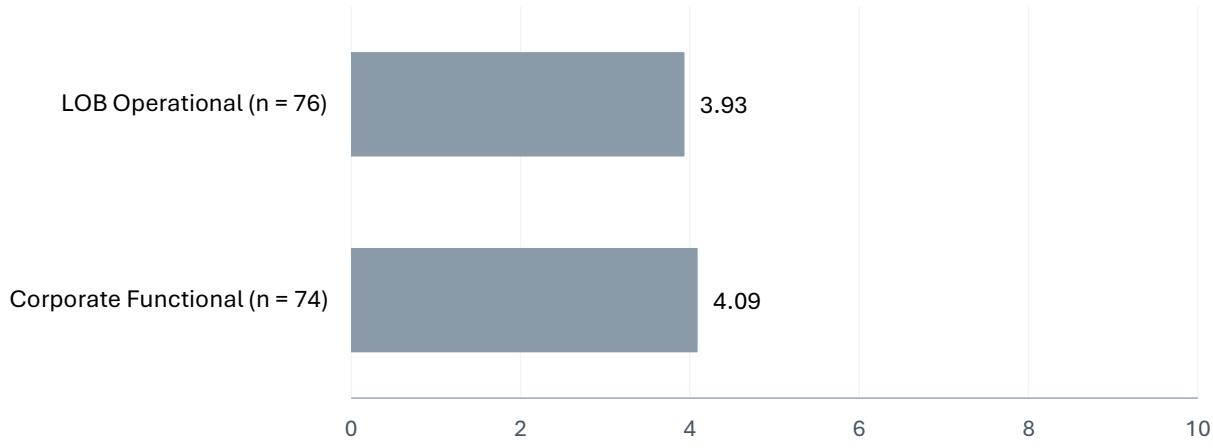
3.6 Class of Work

Segment · 2 classes of work · Full analysis in Part 1, Structural Attributes and Outcome Associations, section 3.2

The two classes of work showed no clearly detected adjusted difference in governance depth. Reported control weaknesses thus extended to both line-of-business operations and corporate functions. This comparison concerns the broad class of work, while the governance analysis continues to examine the specific responsibilities, evidence, and review arrangements within each engagement.

Figure 3.15. The score by class of work, engagements with 10 or more

The two classes of work showed no clearly detected adjusted difference in governance depth.



LOB Operational engagements recorded a score of 3.93 (n = 76 engagements) and Corporate Functional engagements 4.08 (n = 74), with Holm p = 1.00. Distressed governance ratings per respondent stood at 19.79 and 17.91, and written instances per respondent at 3.85 and 3.97, each with Holm p = 1.00.

The two classes differed by less than half a point in every failure category. The widest gap lay in AI technology, at 4.42 for LOB Operational against 4.77 for Corporate Functional engagements, and financial institution readiness stood at 3.55 and 3.57, each with Holm p = 1.00.

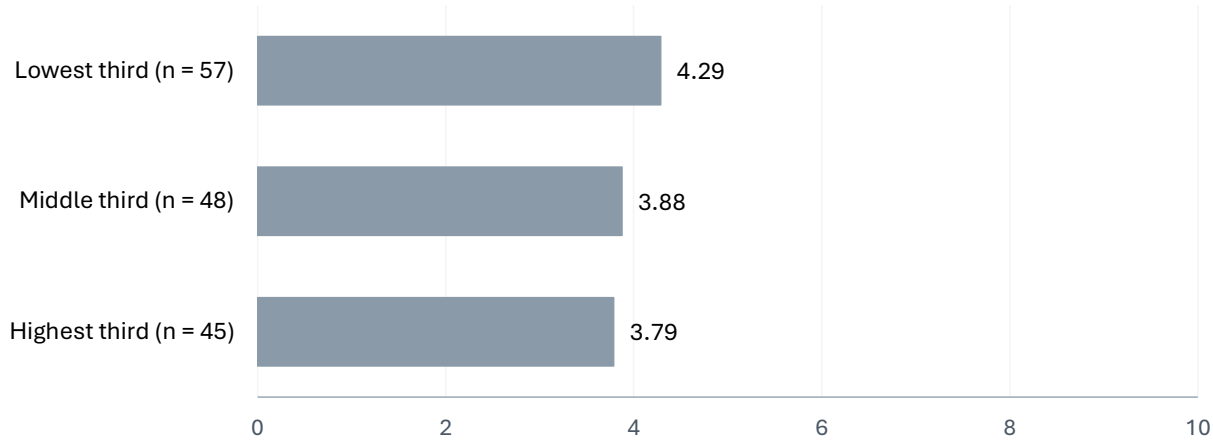
Within each class, the score was associated with cancellation at r = −0.69 on LOB Operational engagements, 30 of them canceled, and at r = −0.75 on Corporate Functional engagements, 26 of them canceled.

3.7 Delivery Team Size

Segment · Delivery team size, current headcount · Full analysis in Part 1, Structural Attributes and Outcome Associations, section 6.1

Delivery team headcount had a small negative observed association with the score, and the conventional correlation test did not detect a clear relationship. The finding provides little support for a simple explanation in which larger teams systematically govern better or worse. Equivalence is a separate question: the study's test uses a correlation margin of ±0.30. The bounded conclusion applies to that margin and this sample.

Figure 3.16. The score by third of delivery team size (current headcount), engagement level



The team-size correlation with the score was r = −0.11 (95% CI −0.26 to 0.06, p = 0.199, n = 150 engagements). The equivalence p value was 0.007 under the study's ±0.30 margin. This bounds the association under that criterion rather than demonstrating identical governance at every team size. Correlations with low-rating counts and written instances per respondent were 0.09 (p = 0.250) and 0.06 (p = 0.461).

3.8 Respondent Role

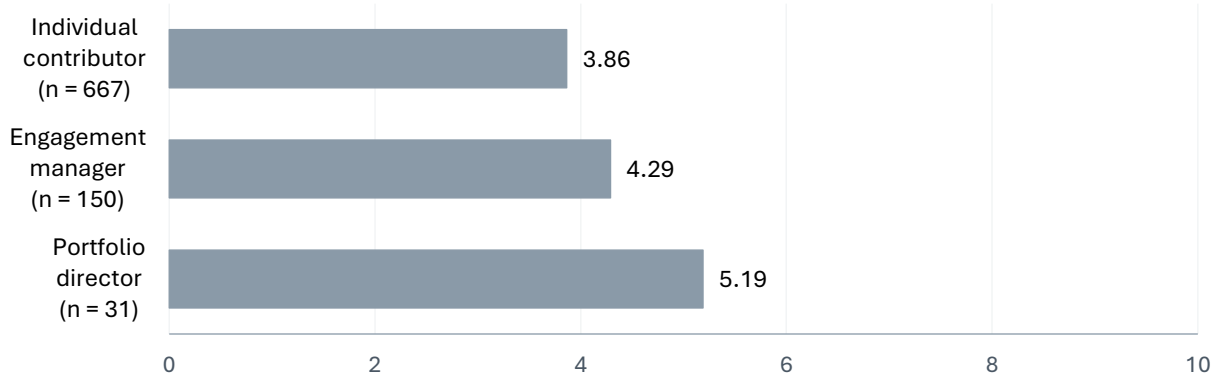
Segment · 3 respondent roles · Full analysis in Part 4, Composition and Experience of the Delivery Workforce, section 1

Governance assessments differed by respondent role, with individual contributors reporting lower scores than engagement managers and portfolio directors. The difference matters because the staff performing agent-related work may see control limitations that higher-level assessments capture less strongly. That explanation remains interpretive. The study compares reported perspectives and does not independently verify which role provides the most accurate assessment.

Figure 3.17. The score by respondent role, respondents with 10 or more

Governance assessments differed by respondent role, with individual contributors reporting lower scores than engagement managers and portfolio directors.

Respondent role
Mean governance, oversight, and control score



Note: n gives the numeric-answer base for each group. These are descriptive comparisons of reported answers.

Individual contributors recorded a score of 3.86 (n = 667 respondents), with 35.83% in the distressed range. Engagement managers recorded 4.29 (n = 150), with 28.00% distressed, and portfolio directors 5.19 (n = 31), with 6.45% distressed (Holm $p < 0.001$). The size of the difference was small, with ω^2 of 0.025.

Distressed governance ratings followed the same order, at 19.73 per individual contributor, 17.56 per engagement manager, and 9.03 per portfolio director (Holm $p < 0.001$). Written instances numbered 4.09, 3.33, and 3.26 per respondent (Holm $p < 0.001$). Every failure category stood lowest among individual contributors and highest among portfolio directors, with co-management at 3.52 against 4.98 and financial institution readiness at 3.41 against

4.66, and the differences held after Holm adjustment for all eight measures (Holm p of 0.002 or less).

The roles differed in the depth of failure they reported, while every governance indicator associated with the score within every role. Each of the 47 governance indicators was associated with the score computed without it within each role, at r = 0.74 to 0.88 among individual contributors, 0.71 to 0.89 among engagement managers, and 0.51 to 0.90 among portfolio directors. The lowest value, 0.51, belonged to the governance of the financial institution's data among the 31 portfolio directors.

The role gradient reached manual alteration. Individual contributors were 418 of the 421 respondents who altered financial-institution-provided data, 481 of the 490 who altered processes, and 276 of the 278 who altered AI outputs, while reporting alteration reached 86 of the 150 engagement managers. Within each role, the score was associated with cancellation at r = −0.64 among individual contributors, −0.63 among engagement managers, and −0.57 among portfolio directors. The two customer management executives are reported as a count only. One individual contributor on a reconciliation engagement reported that the breaks the contributor cleared went unreviewed above the role.

> *“these ask about the two firms together, on my side it is only me and the custodian file. thus i mark low, nobody above me looks at the breaks i clear”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Trade, settlement and reconciliation

4 Governance by Lifecycle Stage, Engagement Age, and Expected Cancellation

Lower scores were associated with canceled engagements and with greater expected cancellation among active engagements. The three active lifecycle stages showed relatively similar mean scores, and engagement age did not show a clear conventional linear association. Together, these comparisons distinguish engagement state from elapsed time. They do not establish a trajectory in which the same engagement's governance improves or deteriorates as it moves between stages.

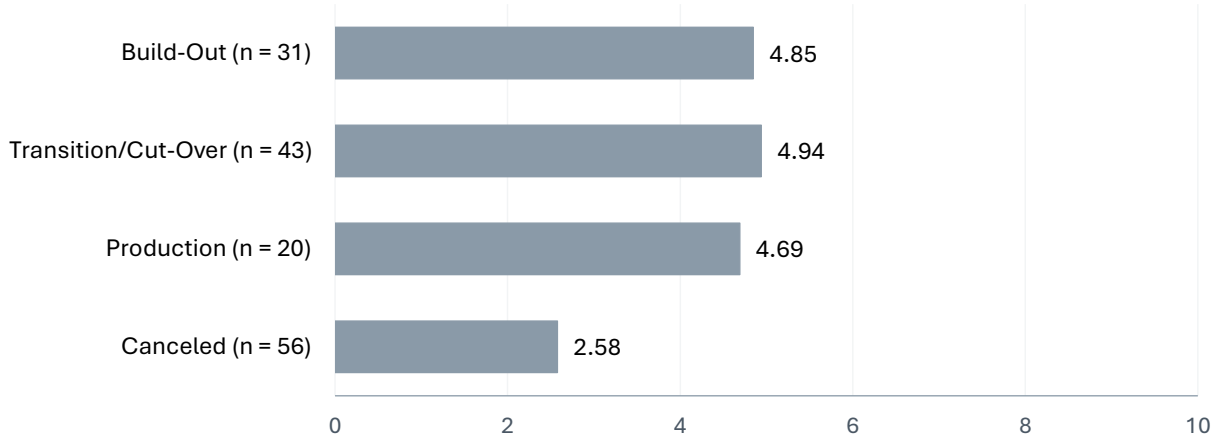
4.1 Lifecycle Stage

Segment · 3 active lifecycle stages and canceled engagements · Full analysis in Part 1, The Immaturity of Agentic AI Outsourcing, section 1.1

Active engagements at all three lifecycle stages had higher mean scores than canceled engagements. The active-stage means were relatively close, while low ratings remained present in Build-Out, Transition/Cut-Over, and Production. This indicates that reported control weaknesses were not confined to early implementation. The comparison is cross-sectional and does not show that a given engagement improved, deteriorated, or changed when it was canceled.

Figure 3.18. The score by engagement status and lifecycle stage, engagements with 10 or more

Active engagements at all three lifecycle stages had higher mean scores than canceled engagements.



The score stood at 4.85 in Build-Out (n = 31 engagements), 4.94 in Transition/Cut-Over (n = 43), and 4.69 in Production (n = 20), against 2.58 on the 56 canceled engagements. Distressed governance ratings rose slightly across the active stages, from 57.39 per engagement in Build-Out to 62.37 in Transition/Cut-Over and 65.95 in Production, and stood at 183.82 per canceled engagement.

Per respondent, distressed governance ratings numbered 10.37, 10.67, and 12.47 across the active stages and 32.14 on canceled engagements. Written instances numbered 3.29, 3.35, and 3.33 per respondent across the active stages, and 4.90 on canceled engagements. The differences across the four levels held after Holm adjustment for all eight measures (Holm $p < 0.001$), while the three active stages stood within 0.25 of one another on the score.

Every failure category stood lowest on the canceled engagements. Co-management fell to 2.19 there against 4.05 to 4.66 in the active stages, and the third-party ecosystem to 3.07 against 5.42 to 5.66. The association of the technology controls to the score varied across the active stages, a pattern set out under AI Technology. An engagement manager on a production engagement reported that neither firm could say who owned a model that gave two figures on two runs, while stopping it quickly worked.

> *“Last month attrition model re-runs twice and gives two different figures. And nobody on either side can say who owns that. We can stop it quickly, that part works.”*
>
> **ENGAGEMENT MANAGER**
> KPO, Corporate and back-office functions

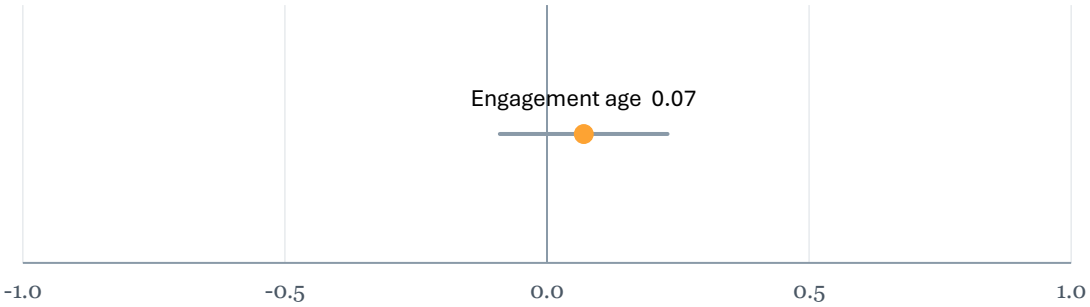
4.2 Engagement Age

Segment · Engagement age in months · Full analysis in Part 1, The Immaturity of Agentic AI Outsourcing, section 1.2

Engagement age had a small observed association with the score and no clearly detected linear relationship in the conventional correlation test. The age comparison offers limited evidence that time in operation alone distinguishes the reported control environment. The recorded equivalence result concerns the specified correlation margin. Neither result measures how the same engagement's governance developed over time.

Figure 3.19. Engagement age: continuous association and descriptive means by ranked third

The conventional correlation test did not clearly detect a linear relationship



r = 0.07; 95% CI −0.09 to 0.23; p = 0.373; n = 150 engagements.

Ranked third of engagement age	Engagements	Mean score
Lowest third	50	3.7999
Middle third	53	4.1724
Highest third	47	4.0364

The thirds are descriptive age groups. Numerical month cutpoints are not reported in the existing exhibit. The continuous estimate above is reproduced from the adjacent analysis; it is not inferred from the grouped means.

Engagement age correlated with the score at r = 0.07 (95% CI −0.09 to 0.23, p = 0.373, n = 150 engagements). The equivalence p value under the recorded ±0.30 margin is 0.002. Correlations with low ratings and written instances per respondent were −0.06 (p = 0.450) and −0.14 (p = 0.085). The findings concern differences among engagements of different ages. They do not show whether governance matured or deteriorated within an engagement.

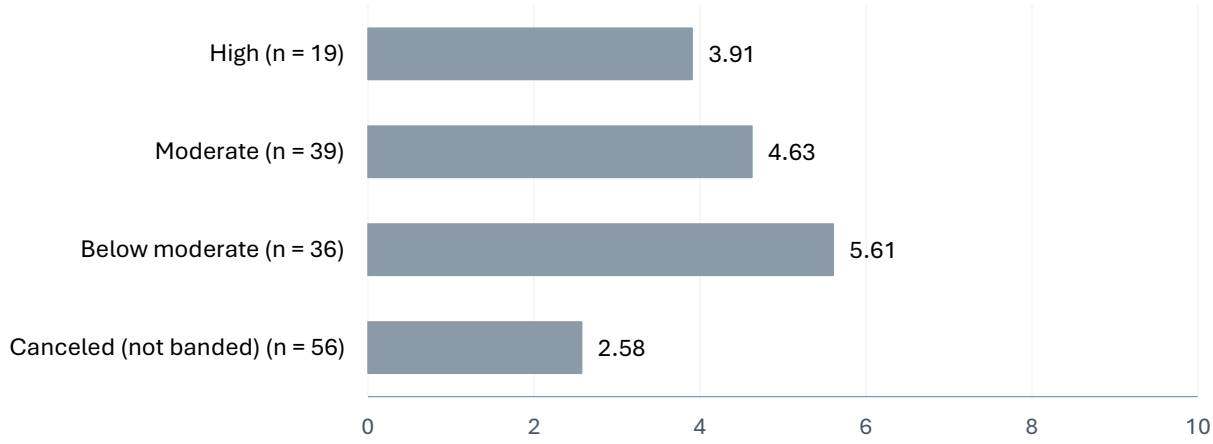
4.3 Expected Cancellation

Segment · 3 bands of cancellation likelihood and canceled engagements · Full analysis in Part 1, Cancellation and the Risk of Cancellation, section 2.1

Among active engagements, higher expected cancellation was associated with lower mean scores. The pattern connects perceived engagement instability with reported governance weakness before observed cancellation is used as a separate comparison. Expected cancellation is a delivery-team assessment, so its association may reflect shared perceptions of the engagement. It is not a validated forecast or evidence that the controls caused the anticipated outcome.

Figure 3.20. The score by cancellation risk band, engagements with 10 or more

Among active engagements, higher expected cancellation was associated with lower mean scores.



The score stood at 5.61 on the 36 active engagements below moderate risk, 4.63 on the 39 at moderate risk, and 3.91 on the 19 at high risk, against 2.58 on the 56 canceled engagements (Holm $p < 0.001$).

Distressed governance ratings numbered 5.67 per respondent below moderate risk, 12.19 at moderate risk, and 18.42 at high risk, or 32.69, 70.69, and 97.16 per engagement. Written instances numbered 2.99, 3.34, and 3.92 per respondent, or 16.89, 19.21, and 21.05 per engagement. Each measure differed across the bands after Holm adjustment (Holm $p < 0.001$).

The gradient ran through all five failure categories. From the high-risk band to the band below moderate risk, co-management rose from 3.59 to 5.32, financial institution readiness from 3.24 to 4.92, vendor capability from 3.48 to 5.01, AI technology from 4.71 to 6.07, and the third-

party ecosystem from 4.48 to 6.38, each with Holm $p < 0.001$. The high-risk engagements had not been canceled, yet their governance stood closer to the canceled engagements than to the active engagements below moderate risk.

“We can stop an agent in minutes, that part is solved and both sides know the number to call. What is not solved is who pays when it has already sent two hundred wrong merchant replies before we stopped it. i would want one line in the operating manual saying which firm owns remediation, decided in advance and not argued after. Also give us a shared change window, right now their release calendar and ours are separate documents.”

ENGAGEMENT MANAGER

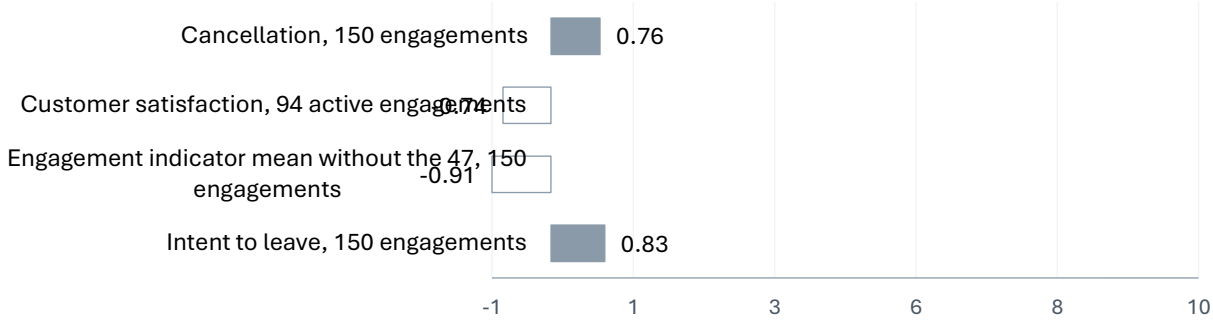
CXM, Payments, cards and disputes

5 Governance Against the Four Study Measures

The score was strongly related to several broader assessments of the engagement. Canceled engagements and respondents intending to leave had lower scores, while customer satisfaction and the mean of the other indicators were positively associated with it. These relationships connect the reported control environment with performance and workforce experience. Their common source, retrospective design, and model dependence limit claims about an independent causal contribution.

Table 3.7. The three measures of reported governance weakness against the four study measures

Distressed governance ratings, r



Study measure	The score, r (95% CI)	After adjustment for the comparison mean (p)	Distressed governance ratings, r
Cancellation, 150 engagements	−0.72 (−0.79 to −0.63)	−0.25 (p = 0.002)	0.76
Customer satisfaction, 94 active engagements	0.80 (0.71 to 0.86)	0.19 (p = 0.069)	−0.74
Engagement indicator mean without the 47, 150 engagements	0.94 (0.92 to 0.95)	Not tested	−0.91
Intent to leave, 150 engagements	−0.84 (−0.88 to −0.79)	Not tested	0.83

Study measure	After adjustment for the comparison mean (p)	Written instances, r	After adjustment for the comparison mean (p)
Cancellation, 150 engagements	0.43 (p < 0.001)	0.64	0.22 (p = 0.007)
Customer satisfaction, 94 active engagements	−0.16 (p = 0.130)	−0.55	−0.07 (p = 0.516)
Engagement indicator mean without the 47, 150 engagements	Not tested	−0.77	Not tested
Intent to leave, 150 engagements	Not tested	0.70	Not tested

Note: Correlations are at the engagement level, with the paired base stated for each result. The adjusted comparison uses the mean of the 53 indicators outside the governance, oversight, and control score. It does not apply when that comparison mean is itself the variable being examined, and no adjusted intent-to-leave test is reported here. All cancellation discrimination is retrospective and within sample. AUC ranges from 0.5 for chance ordering to 1 for complete ordering in the stated score direction. It is not evidence of validated prospective prediction.

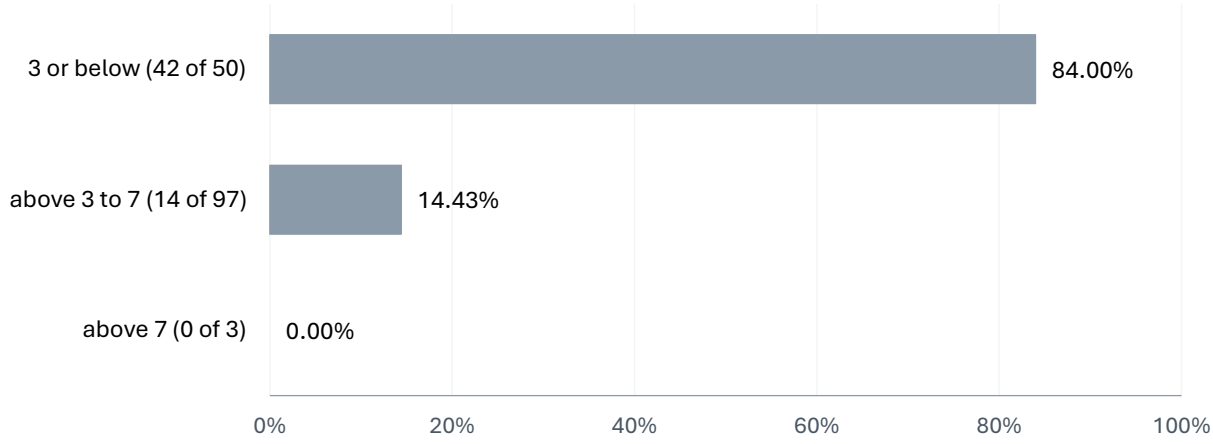
5.1 Cancellation

Study measure · Cancellation · Full analysis in Part 1, Cancellation and the Risk of Cancellation, section 1.1

Already canceled engagements had substantially lower mean scores than active engagements, and a negative association remained after adjustment for the other-indicator mean. This makes cancellation an important descriptive contrast for understanding the reported control environment. It does not establish whether governance weaknesses preceded cancellation, followed it, or were assessed differently after it. The discrimination result describes separation within this sample.

Figure 3.21. Share of engagements canceled, by numeric score band

Cancellation was concentrated among engagements with low scores



The 56 canceled engagements recorded a score of 2.58, against 4.86 on the 94 active engagements (Welch $p < 0.001$), and the score separated the two with an area under the curve of 0.95. It was associated with cancellation at $r = -0.72$ (95% CI −0.79 to −0.63, $n = 150$ engagements). After adjustment for the comparison mean, the association remained at partial $r = -0.25$ ($p = 0.002$).

The two counts kept larger ties than the score. Distressed governance ratings numbered 32.14 per respondent on canceled engagements against 10.95 on active ones, an association with cancellation of $r = 0.76$ that remained at partial $r = 0.43$ after that adjustment ($p < 0.001$). Written instances numbered 4.90 per respondent against 3.33, an association of $r = 0.64$ that remained at partial $r = 0.22$ ($p = 0.007$).

Every failure category carried the gap. On canceled engagements co-management stood at 2.19 against 4.05 to 4.66 across the active stages, financial institution readiness at 2.38 against 4.19 to 4.36, vendor capability at 2.44 against 4.25 to 4.76, AI technology at 3.14 against 5.37 to 5.54, and the third-party ecosystem at 3.07 against 5.42 to 5.66. Within each role, the score was associated with cancellation at r = −0.64 among individual contributors, −0.63 among engagement managers, and −0.57 among portfolio directors.

The 56 canceled engagements had been canceled when their delivery teams answered, so the contrast is descriptive. On one canceled engagement, an audit-readiness report marked a reporting rule as documented whenever any version of the rule existed, and the gap surfaced only when the auditor asked.

> *"The vendor's audit-readiness dashboard marked a reporting rule as fully documented once any version of the rule existed in the knowledge base, without checking whether that version matched the rule the agent was actually executing, so undocumented rule changes never surfaced as a gap until the auditor asked directly."*
>
> **INDIVIDUAL CONTRIBUTOR**
> KPO, Regulatory, audit, tax and legal

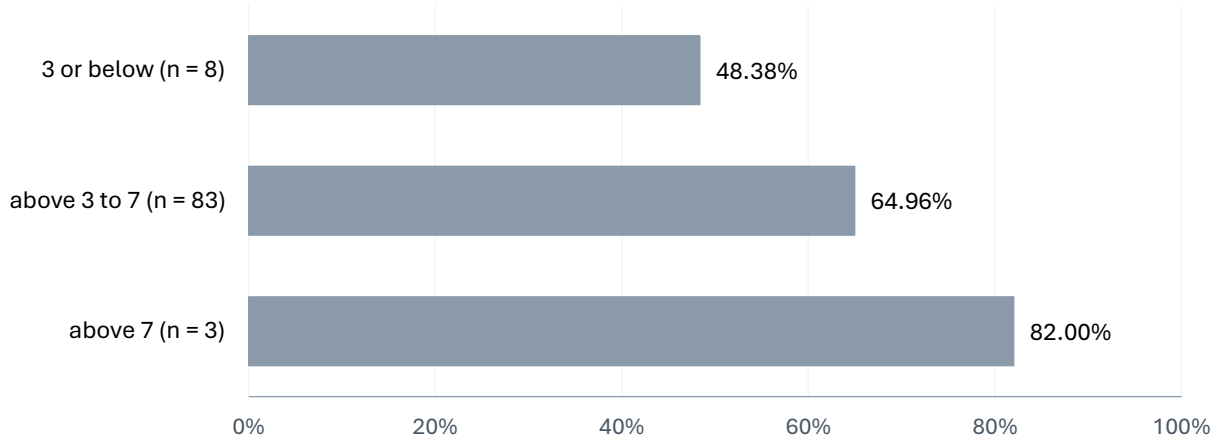
5.2 Customer Satisfaction

Study measure · Customer satisfaction · Full analysis in Part 1, Customer Satisfaction, Delivery Performance, and Measured Return, section 1.1

Higher scores were strongly associated with higher customer satisfaction among active engagements. Much of that association was shared with the engagement's standing on the other indicators, and the adjusted estimate was smaller and did not meet the nominal significance threshold. The finding connects reported control quality with the institution's assessment of delivery while limiting a claim that governance contributes a statistically distinct relationship under this model.

Figure 3.22. Customer satisfaction, mean percent, by numeric score band, 94 active engagements

Higher scores were strongly associated with higher customer satisfaction among active engagements.



Note: partial r of the score with customer satisfaction, after adjustment for the mean of the other 53 indicators: 0.19 (p = 0.069, against r = 0.80 before the control, 94 active engagements).

Among the 94 active engagements, the score correlated with customer satisfaction at r = 0.80 (95% CI 0.71–0.86). After adjustment for the mean of the other 53 indicators, partial r was 0.19 (p = 0.069). The smaller estimate did not meet the nominal threshold, with no inference that the true adjusted association is zero.

Low-rating counts and written-instance counts also correlated with customer satisfaction. The respective raw correlations were −0.74 and −0.55, falling to partial r = −0.16 (p = 0.130) and −0.07 (p = 0.516) after adjustment. Neither adjusted comparison met the nominal threshold. Restriction to active engagements removes the already-canceled group, but the remaining shared-source and conditional-model limitations still apply.

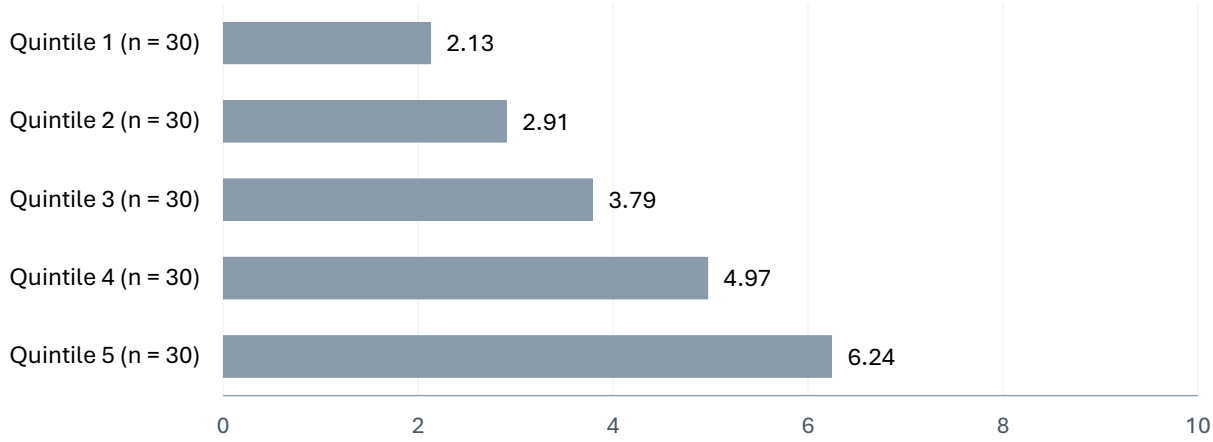
5.3 Engagement Indicator Mean and Indicator Quintile

Study measure · Engagement indicator mean and indicator quintile · Full analysis in Part 1, The Engagement Indicator Mean and the Interdependence of the Four Measures, section 1

The score was strongly correlated with the mean of the 53 indicators outside it. This connects reported governance quality with the wider assessment of engagement performance and capability. The two measures use disjoint indicator sets, so the association is empirical rather than a literal overlap of items. Adjustment asks whether a relationship remains after accounting for that wider standing, with the limitations of the composite comparison stated explicitly.

Figure 3.23. The score by engagement indicator mean quintile, engagements with 10 or more

The score was strongly correlated with the mean of the 53 indicators outside it.



The score correlated with the mean of the other 53 indicators at r = 0.94 (95% CI 0.92–0.95, n = 150 engagements). The strong empirical relationship explains why the adjusted comparisons are informative but also difficult to interpret. Accounting for a related composite can substantially alter a coefficient, and category composition affects what the adjustment removes.

Written instances fell from 29.27 per engagement in the lowest indicator quintile to 26.10, 21.47, 19.87, and 14.30 in the four quintiles above it, each quintile holding 30 engagements (Welch F = 57.16, Holm p < 0.001). Distressed governance ratings fell more steeply, from 201.63 per engagement in the lowest quintile to 167.17, 109.80, 45.17, and 12.03 (Welch F = 550.56, Holm p < 0.001). The highest quintile still carried written instances on all 30 of its engagements.

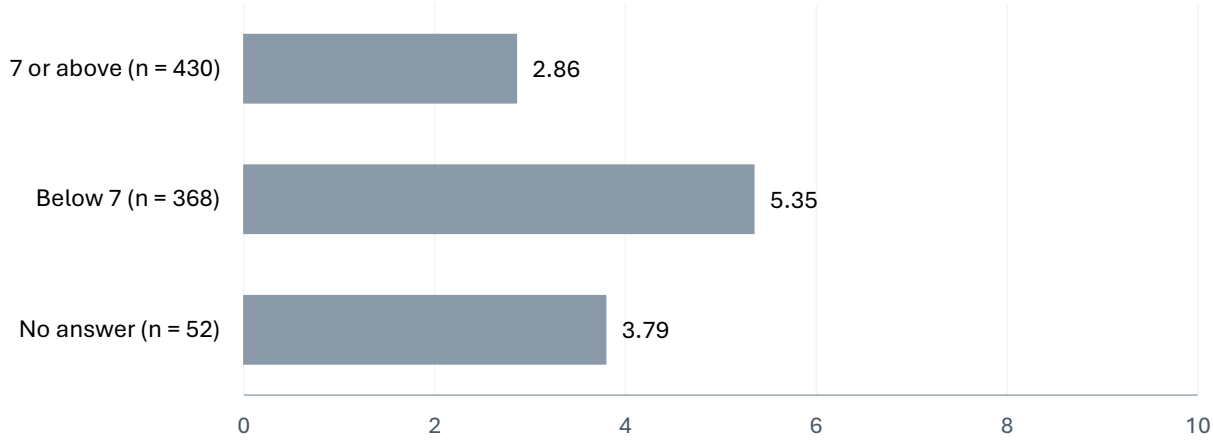
5.4 Intent to Leave

Study measure · Intent to leave · Full analysis in Part 4, Intent to Leave, section 1.1

Respondents reporting stronger intent to leave also reported substantially lower scores. This connects the control environment with the experience of the delivery workforce, particularly where staff describe manual correction, unresolved issues, and work that remains poorly visible to oversight. The comparison concerns stated intention, not subsequent departure. It also shares a respondent source with the score and does not identify the direction of the relationship.

Figure 3.24. The score by intent to leave, respondents with 10 or more

Respondents reporting stronger intent to leave also reported substantially lower scores.



Scores were 3 or below for 59.53% of the higher-intention group and 2.99% of the lower-intention group. The 430 respondents who rated their intent to leave at 7 or above stood at 2.86, the 368 below 7 at 5.35, and the 52 who gave no answer at 3.79 (Holm $p < 0.001$). Distressed governance ratings numbered 29.15 per respondent at 7 or above against 6.77 below 7, and written instances 4.71 against 2.99 (Holm $p < 0.001$).

Every failure category followed the same order. Co-management stood at 2.48 among respondents intending to leave against 5.04 among the rest, and the third-party ecosystem at 3.39 against 6.06. At the engagement level, the score was associated with intent to leave at $r = -0.84$ (95% CI −0.88 to −0.79, n = 150 engagements), and at the respondent level at $r = -0.78$ (cluster-robust $p < 0.001$, n = 798 respondents).

Intent to leave did not mark the lowest the score of the study. The canceled engagements stood lower, at 2.58, and so did the respondents who rated their own understanding of the

governing obligations 1 to 3, at 2.41. Among the challenges that reduced their job satisfaction, respondents at intent to leave 7 or above named reported governance weaknesss directly. One individual contributor on a canceled engagement, who rated intent to leave at 10, named three.

Co-Management and Governance Across the Other Four Categories

Co-management ratings consistently connected shared operating arrangements with governance in the other four categories. The 20 indicators concern authority, review, change, evidence, and response across the financial institution and vendor. Their low-rating share was 49.26%, based on 7,866 of 15,968 ratings. The governance, oversight, and control score provides a common outcome for examining these responsibilities, while the adjusted analysis remains conditional on its unequal category composition.

All 20 co-management indicators were strongly associated with the score after the tested indicator was omitted, with engagement-level r = 0.90–0.93. Correlations with the other four categories were 0.75–0.80. Under the main adjustment, partial r ranged from 0.21 to 0.31. All 20 met nominal $p < 0.05$ and survived Holm adjustment when considered as one 20-test family. The inference changes with the family: only two of the 20 survive Holm adjustment across all 47 cross-category tests. Subgroup results reuse overlapping respondents and engagements and are not independent replications.

The adjusted comparison has an important asymmetry. All 20 co-management indicators are in the score and none is in the 53-item comparison mean. That mean instead contains 18 other vendor capability indicators, 13 readiness indicators, 12 AI technology indicators, and 10 third-party indicators. Removing an item's category from the score also leaves different numbers of items in the outcome. The results identify associations conditional on this construction. They do not establish that co-management uniquely causes weaknesses in the other categories.

Among respondents who rated a co-management indicator 1 to 3, 61.19% to 65.21% had a score of 3 or below, against 2.65% to 5.30% among those who rated it 4 to 7. None of the respondents who rated a co-management indicator 8 to 10 had a score of 3 or below. Table 3.8 sets out the 20 indicators against these measures.

Table 3.8. The 20 co-management indicators against reported governance weakness across the engagement

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Co-Management of the Engagement	249 of 385 (64.68%)	0.91 (0.88 to 0.93)	0.76
Decision Rights Over Agent Action Without Human Approval	257 of 406 (63.30%)	0.91 (0.88 to 0.94)	0.77

Table 3.8. The 20 co-management indicators against reported governance weakness across the engagement (continued)

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Ownership of Each Workflow Step	253 of 388 (65.21%)	0.92 (0.89 to 0.94)	0.78
Accountability for Problems Arising From an Agent's Action	246 of 402 (61.19%)	0.93 (0.90 to 0.95)	0.80
A Single Shared View of Agent Risk	245 of 383 (63.97%)	0.92 (0.89 to 0.94)	0.78
Co-Management of Issues Arising From Agent Work	249 of 391 (63.68%)	0.92 (0.89 to 0.94)	0.78
Resolution Speed Against the Spread of Agent-Caused Problems	248 of 400 (62.00%)	0.91 (0.88 to 0.94)	0.77
Speed of Stopping a Misbehaving Agent in Live Operations	261 of 406 (64.29%)	0.92 (0.89 to 0.94)	0.79
Joint Change Control Against the Speed of Agent Behavior Change	249 of 399 (62.41%)	0.92 (0.89 to 0.94)	0.79
Visibility of Agent Updates to the Other Firm Before They Take Effect	250 of 394 (63.45%)	0.91 (0.88 to 0.94)	0.77
Joint Governance of Agent Behavior Change Between Formal Releases	253 of 388 (65.21%)	0.90 (0.87 to 0.93)	0.76
Placement of Human Review Points	253 of 393 (64.38%)	0.91 (0.88 to 0.94)	0.77
Human Review Capacity Against Agent Output Volume	244 of 381 (64.04%)	0.90 (0.86 to 0.92)	0.75
Genuine Scrutiny at Human Review Points	251 of 393 (63.87%)	0.92 (0.89 to 0.94)	0.78
Reliability of the Joint Quality Assurance Loop	244 of 377 (64.72%)	0.91 (0.88 to 0.94)	0.77
Quality Measures That Capture Agent Errors	250 of 405 (61.73%)	0.93 (0.90 to 0.95)	0.79
Each Firm's Visibility of the Other Firm's Agent Actions	245 of 392 (62.50%)	0.92 (0.89 to 0.94)	0.79
Joint Decision Speed Against Agent Speed	245 of 381 (64.30%)	0.92 (0.89 to 0.94)	0.79
Fit of the Joint Governance Model to Work Performed by Agents	251 of 395 (63.54%)	0.92 (0.89 to 0.94)	0.78

Table 3.8. The 20 co-management indicators against reported governance weakness across the engagement (continued)

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Readiness of the Co-Management Model for Expected Agent Autonomy	254 of 407 (62.41%)	0.93 (0.90 to 0.95)	0.80

Indicator	Adjusted partial r (p)	Negative written answers
Co-Management of the Engagement	0.21 (p = 0.010)	127 of 209
Decision Rights Over Agent Action Without Human Approval	0.25 (p = 0.002)	34 of 48
Ownership of Each Workflow Step	0.25 (p = 0.003)	70 of 82
Accountability for Problems Arising From an Agent's Action	0.22 (p = 0.007)	45 of 51
A Single Shared View of Agent Risk	0.25 (p = 0.002)	19 of 26
Co-Management of Issues Arising From Agent Work	0.26 (p = 0.001)	39 of 54
Resolution Speed Against the Spread of Agent-Caused Problems	0.25 (p = 0.002)	16 of 18
Speed of Stopping a Misbehaving Agent in Live Operations	0.26 (p = 0.001)	59 of 73
Joint Change Control Against the Speed of Agent Behavior Change	0.31 (p < 0.001)	57 of 71
Visibility of Agent Updates to the Other Firm Before They Take Effect	0.25 (p = 0.002)	38 of 48
Joint Governance of Agent Behavior Change Between Formal Releases	0.23 (p = 0.004)	12 of 14
Placement of Human Review Points	0.23 (p = 0.004)	75 of 105
Human Review Capacity Against Agent Output Volume	0.25 (p = 0.003)	134 of 145
Genuine Scrutiny at Human Review Points	0.25 (p = 0.002)	59 of 66
Reliability of the Joint Quality Assurance Loop	0.21 (p = 0.012)	22 of 29
Quality Measures That Capture Agent Errors	0.26 (p = 0.002)	11 of 14
Each Firm's Visibility of the Other Firm's Agent Actions	0.22 (p = 0.007)	62 of 66

Table 3.8. The 20 co-management indicators against reported governance weakness across the engagement (continued)

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Joint Decision Speed Against Agent Speed	0.23 (p = 0.005)	13 of 14	
Fit of the Joint Governance Model to Work Performed by Agents	0.26 (p = 0.001)	9 of 10	
Readiness of the Co-Management Model for Expected Agent Autonomy	0.22 (p = 0.008)	2 of 3	

Note: The score in the rating-band and item-to-score comparisons excludes the tested indicator. Cross-category correlations exclude its entire category and use 150 engagements. The partial correlation accounts for the mean of the 53 indicators outside the score. Cross-category p values are nominal. All 20 pass Holm within a family of 20, whereas two pass within a family of all 47 cross-category comparisons. Chi-square compares rating-band counts, and Cramér's V summarizes the strength of that categorical association from 0 to 1. The ordinary respondent-level chi-square calculations do not adjust for dependence within engagements. Cluster-robust mean tests account for that dependence in their standard errors. Zero observations in a small band are not proof of no risk. Monte Carlo calculation approximates a contingency-test p value through repeated simulated tables when expected counts are sparse. It does not correct engagement clustering.

> *"What a number cannot carry is that the two governance structures never met. We had our review, they had theirs, and no forum existed where the two were reconciled and someone could say who owned a step when an agent had performed it."*
>
> **PORTFOLIO DIRECTOR**
>
> KPO, Corporate and back-office functions

1 Joint Governance and Decision Rights

Co-management, decision rights, step ownership, and accountability each had strong observed relationships with the wider score. Together they define who can decide, who must act, and who owns the response when agent work goes wrong. Their similar statistical patterns support a common concern about shared authority, while each indicator addresses a distinct responsibility boundary. The evidence does not establish that a formal assignment alone makes that authority effective.

> *These questions assume somebody actually decides. From where I sit nothing is decided, it is only carried to next weeks call and then the one after. Cannot put number of 1 to 10 on a thing that never happens"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Corporate and back-office functions

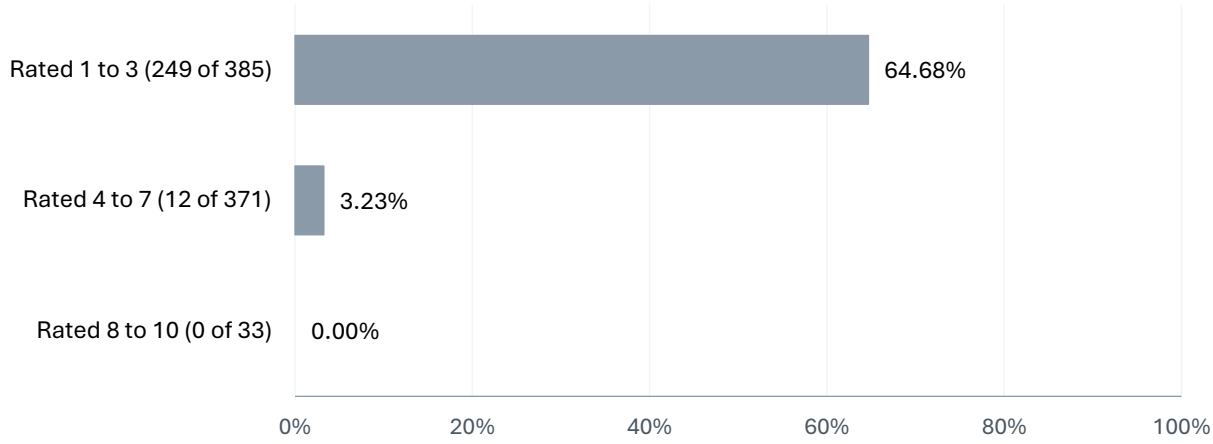
1.1 Co-Management of the Engagement

Indicator C136 · Rate how effectively the customer and vendor jointly manage this engagement now that agents perform part of the work · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 3.6 · Full statistical analysis ›

Weak co-management was associated with lower scores beyond the tested indicator. Co-management needs to coordinate the resolution of the same operational problem across the two firms. Written accounts describe forums that existed but did not settle decisions, alongside accounts of more effective joint working. That contrast makes the operation of coordination, rather than the number of negative comments, the central finding.

Figure 3.25. Co-Management of the Engagement: share with a score of 3 or below by indicator rating.

Weak co-management was associated with lower scores beyond the tested indicator.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Scores were 3 or below for 249 of 385 (64.68%) at indicator ratings 1–3. The corresponding counts were 12 of 371 (3.23%) at indicator ratings 4–7 and 0 of 33 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 339.21, $p < 0.001$, Cramér's V = 0.66.

Co-management was associated with the score at r = 0.91 (95% CI 0.88 to 0.93, n = 150 engagements) and with governance in the other four categories at r = 0.76. After adjustment for the mean of the other 53 indicators, the association remained at partial r = 0.21 (p = 0.010), the second-weakest adjusted association of the 20.

The association was observed within the canceled engagements at r = 0.78 and across the active stages at 0.80 to 0.84, and it held after Holm adjustment in 85 of the 87 segment levels tested. Within role it stood at 0.87 among individual contributors, 0.85 among engagement

managers, and 0.87 among portfolio directors. Among individual contributors alone, 66.88% (206 of 308) of those who rated co-management 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.71 (n = 385), against 5.23 (n = 404) among the rest (cluster-robust $p < 0.001$).

Co-management drew 209 written answers from 199 respondents on 115 engagements, the most of the 20 indicators, and 94 of them came from canceled engagements. Of the 127 negative answers, 83 came from canceled and 44 from active engagements, and 28 (22.05%) described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (17), change discovered only through failure (6), and raising problems seen as futile (5). The negative answers described joint forums that existed on paper and decided nothing. The answers that were not negative described working forums beside a control that still failed, most often the authority to let an agent act without review.

> *"I want to be fair, the score is low because of one thing, escalation. We have a weekly joint call and it is polite and nothing gets decided in it. When the agent started pulling market share numbers from the wrong vintage of a report, I raised it on that call, their side logged it in their tracker, we logged it in ours, and for three weeks each side thought the other one was fixing it. My own management told me to keep it at working level and not put it in the client summary. So there are two versions of the issue list and neither one is complete. That is what the 1s are, not the people."*
>
> **ENGAGEMENT MANAGER**
>
> KPO, Corporate and back-office functions

> *"Where I felt genuine progress is in the joint forum itself. In the first month the two sides came to it with separate issue logs, and we spent most of the hour reconciling numbers rather than deciding anything; that is behind us now, there is one list, both parties own it, and the weekly escalation call rarely runs over. Where there has been no movement is the question of who may authorise an agent to close an alert without a human looking at it. That has been open since the design workshops, it has been discussed about at every steering committee, and each time it is deferred to the next one, largely because neither side wishes to be the party whose name sits against that decision if a hit is missed. Review capacity is the same story in a different form."*
>
> **PORTFOLIO DIRECTOR**
>
> KPO, Financial crime and client onboarding

Part 2 (Co-Management, section 1.1) reports the association of co-management with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 3.6) its association with manual alteration.

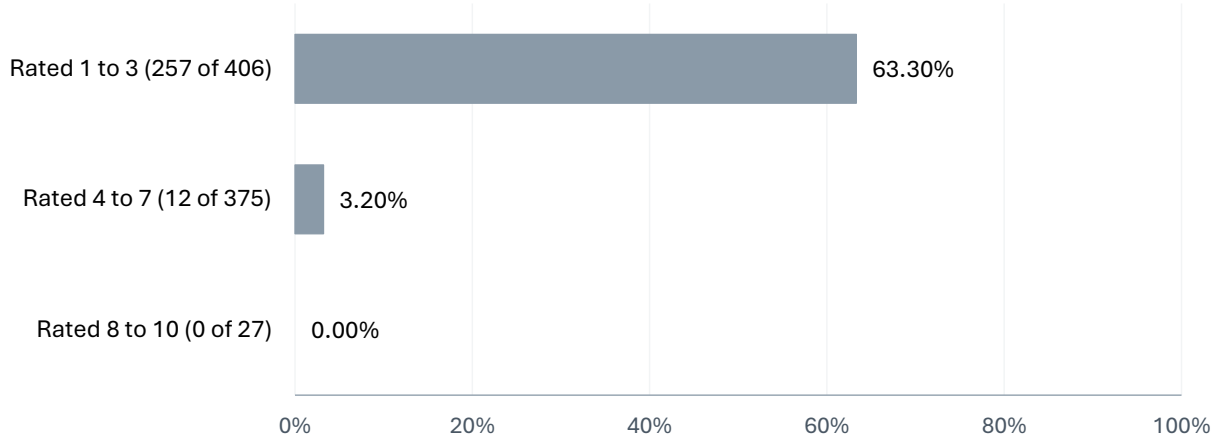
1.2 Decision Rights Over Agent Action Without Human Approval

Indicator C137 · Rate how clearly the two firms have settled who decides what agents may do without human approval · Full statistical analysis ›

Unclear authority for agent action was associated with a weaker reported control environment. Decision rights determine which actions require human approval and which the two firms permit an agent to take independently. Where that boundary remains unsettled, a formal management structure may leave operational authority ambiguous. The ratings establish an association with the score, while the study does not directly observe every authorization decision or unauthorized action.

Figure 3.26. Decision Rights Over Agent Action Without Human Approval: share with a score of 3 or below by indicator rating.

Unclear authority for agent action was associated with a weaker reported control environment.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

The 406 respondents who rated decision rights 1 to 3 included 257 with a score of 3 or below. Among the 375 who rated them 4 to 7, 12 (3.20%) had a score of 3 or below, and none of the 27 who rated them 8 to 10 did (chi-square = 331.00, $p < 0.001$, Cramér's V = 0.64).

Decision rights were associated with the score at $r = 0.91$ (95% CI 0.88 to 0.94, $n = 150$ engagements) and with governance in the other four categories at $r = 0.77$. After adjustment for

the mean of the other 53 indicators, the association remained at partial r = 0.25 (p = 0.002), the sixth-strongest of the 20.

The association was observed within the canceled engagements at r = 0.77 and across the active stages at 0.80 to 0.84, and after Holm adjustment it held in 85 of the 87 segment levels tested. Within role it stood at 0.87 among individual contributors, 0.85 among engagement managers, and 0.86 among portfolio directors. Among individual contributors alone, 65.24% (214 of 328) of those who rated decision rights 1 to 3 had a score of 3 or below. Respondents who rated them distressed recorded a score of 2.74 (n = 406), against 5.30 (n = 402) among the rest (cluster-robust p < 0.001).

Forty-eight respondents on 42 engagements wrote about the limit of agent autonomy, one answer each, and 34 of the answers were negative, 13 from canceled and 21 from active engagements. Five of the 34 described one of the 26 forms of a control bypassed or not working, three of them problems raised with no response. Respondents described reviewers who approved agent work on trust in the absence of a written limit, and edge cases that reached a manager every evening. One individual contributor on a reconciliation engagement set out the missing control in three parts, a named approver on each side, one shared list, and a written limit on what the agent could clear alone. An engagement manager on another reconciliation engagement described what the missing limit cost the team.

> *“i will use this box properly since nobody asked before. what would have to change is one name. one person on their side and one on ours who can say yes to a break over a certain amount, that is all. right now a break comes up, i put it in the shared tracker, their ops person says it needs their treasury, treasury says the custodian must confirm, custodian says ask the client. ten days gone. also by then the next cycle is on top of it and the same break is now three breaks. so second thing, the tracker itself. we have ours and they have theirs and the numbers dont agree, and so half of every call is spent agreeing which list is correct instead of clearing anything. one list. it doesnt need a system, a spreadsheet would do. third, and this is the real one, somebody has to write down what the agent is allowed to clear on its own. under five hundred, matched to the statement line, clear it. we never got that number. so everything comes to a human, and the humans are waiting on each other. i said this in month two, kindly nothing came of it. which is why i am not hopeful, but you asked what would change it and that is what”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Trade, settlement and reconciliation

> *What this costs us is people. My analysts are checking machine output at nine at night, work a tool was supposed to remove, because neither side will put a name against what the agent may clear on its own. Two of my better ones have asked to move. I have nothing to offer them to stay."*
>
> **ENGAGEMENT MANAGER**
>
> BPO, Trade, settlement and reconciliation

Part 2 (Co-Management, section 1.2) reports the association of decision rights with the four study measures.

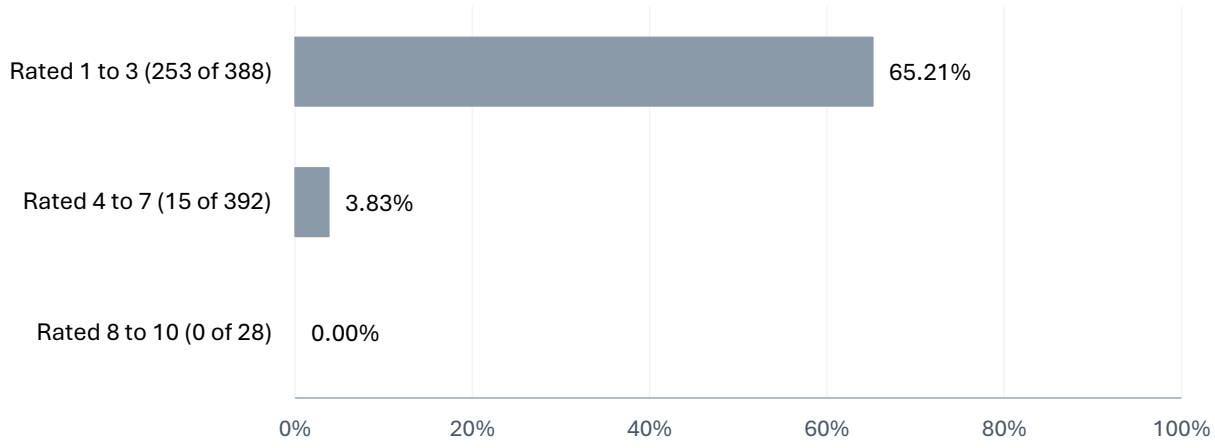
1.3 Ownership of Each Workflow Step

Indicator C138 · Rate how clearly ownership of each workflow step is assigned when agents, vendor staff, and customer staff work in the same process · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 3.7 · Full statistical analysis ›

Unclear ownership of workflow steps was associated with low scores. This concerns the point at which agents, delivery team members, and financial institution staff share a process but responsibility for a particular step is not settled. The governance issue is the assignment of responsibility across the workflow, distinct from accountability after an error. Similar score relationships across these controls do not make those responsibilities interchangeable.

Figure 3.27. Ownership of Each Workflow Step: share with a score of 3 or below by indicator rating.

Unclear ownership of workflow steps was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 388 respondents who rated step ownership 1 to 3, 253 had a score of 3 or below. Among the 392 who rated it 4 to 7, 15 (3.83%) had a score of 3 or below, and none of the 28 who rated it 8 to 10 did (chi-square = 345.81, p < 0.001, Cramér's V = 0.65).

Step ownership was associated with the score at r = 0.92 (95% CI 0.89 to 0.94, n = 150 engagements) and with governance in the other four categories at r = 0.78. After adjustment for the mean of the other 53 indicators, the association remained at partial r = 0.25 (p = 0.003).

The association was observed within the canceled engagements at r = 0.74, the joint weakest of the 20 together with genuine scrutiny at review points, and it ranged across the active-stage groups from 0.80 in Build-Out to 0.85 in Production. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.87 among individual contributors, 0.86 among engagement managers, and 0.82 among portfolio directors. Among individual contributors alone, 66.67% (212 of 318) of those who rated step ownership 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.68 (n = 388), against 5.22 (n = 420) among the rest (cluster-robust p < 0.001).

Step ownership drew 82 written answers from 81 respondents on 58 engagements, and 70 of them were negative, 39 from canceled and 31 from active engagements. Of the 70, 21 (30.00%) described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (11), change discovered only through failure (6), and raising problems seen as futile (4). Respondents described errors that passed between the two firms with each assuming the other had checked, and decisions that waited for weeks while each firm assigned them to the other. On one canceled engagement the team spent an hour on a call establishing which firm owned an error and then corrected it itself.

> *the position is like this. there is no single owner of the allocation steps. my team prepares the expense split, the tool proposes the allocation, our lead looks at it, then it goes across to the fund controller side, and nowhere is it written which of those is the decision. so in practice the last person to touch the file is the decision. second thing, the tool behaves differently after an update and there is no joint forum where that is tabled. we find out because the output looks different that morning. i mention it in the daily call, that is the whole of the process. third, nobody has told us what the tool is permitted to do without a person approving. so i assume nothing, and i open each line myself, do note this roughly doubles the time on close days. the review on their side is a sign off and not a review. they sign what we send, and i am not saying they are wrong for that, they cannot check what they cannot see from their end. and when a query does come back it is three weeks later and the person who did the work is on another fund by then. one more, when we change something on our side there is no way for them to see it before it takes effect, and the same in reverse. that is the picture from where i sit"*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Trade, settlement and reconciliation

> *honestly my team is losing patience. when something go wrong nobody say which side own it, so we sit in call for hour and then still we fix it ourself. two of my colleague already ask for transfer"*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Trade, settlement and reconciliation

Part 2 (Co-Management, section 1.3) reports the association of step ownership with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 3.7) its association with manual alteration.

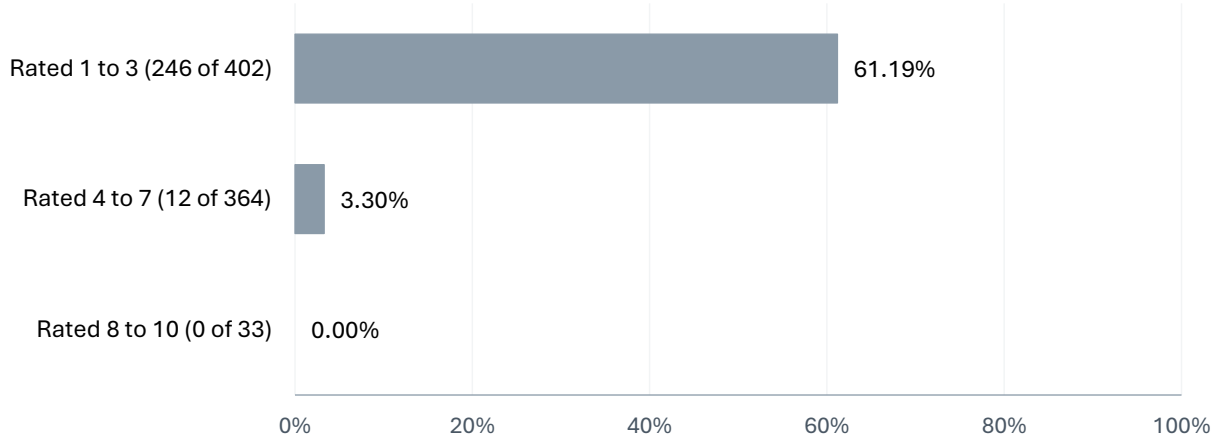
1.4 Accountability for Problems Arising From an Agent's Action

Indicator C139 · Rate how clearly accountability lands on one firm when an agent's action causes a problem in this engagement · Full statistical analysis ›

Accountability for agent error was strongly associated with the wider score. Its control function is to establish which firm owns the response when an agent's action causes a problem. This differs from assigning routine workflow steps: an engagement may have detailed task ownership and still leave remediation responsibility unresolved. The overall association and the low-score share within a rating band describe different aspects of that relationship.

Figure 3.28. Accountability for Problems Arising From an Agent's Action: share with a score of 3 or below by indicator rating.

Accountability for agent error was strongly associated with the wider score.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

The indicator was associated with the score at r = 0.93 (95% CI 0.90 to 0.95, n = 150 engagements), second among the 47 governance indicators to the readiness of the joint model for expected agent autonomy. Scores were 3 or below for 246 of 402 (61.19%) at indicator ratings 1–3. The corresponding counts were 12 of 364 (3.30%) at indicator ratings 4–7 and 0 of 33 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 309.30, $p < 0.001$, Cramér's V = 0.62.

The strong overall association and the conditional low-score share answer different questions. Accountability correlated with governance in the other four categories at r = 0.80. After adjustment for the comparison mean, partial r was 0.22 (nominal p = 0.007). These results support a model-qualified relationship with the wider control environment, without a tested ranking of accountability's importance against other controls.

The association was observed within the canceled engagements at r = 0.76 and across the active stages at 0.82 to 0.88, and after Holm adjustment it held in 86 of the 87 segment levels tested. Within role it stood at 0.87 among individual contributors, 0.88 among engagement managers, and 0.89 among portfolio directors. Among individual contributors alone, 62.58% (204 of 326) of those who rated accountability 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.76 (n = 402), against 5.35 (n = 397) among the rest (cluster-robust p < 0.001).

Accountability drew 51 written answers from 50 respondents on 37 engagements, and 45 were negative, 20 from canceled and 25 from active engagements. Of the 45, 11 (24.44%) described one of the 26 forms of a control bypassed or not working, most often unclear liability or regulatory accountability (5), change discovered only through failure (3), and problems raised with no response (2). Respondents described errors that passed through three parties, each of which had done what was asked. An engagement manager on a payments engagement where an agent could be stopped within minutes asked for one line, settled in advance, naming the firm that owned remediation.

> *“I want to spend the whole of this box on one point, because the rest of it is ordinary. Ownership of a fee correction is, I would say, still unassigned. The agent proposes the adjustment, our analyst validates it, the bank's billing supervisor approves it, and on paper that is three clean steps. In practice, when a correction is wrong, all three of those parties can point at the other two and each of them is correct. I have run four engagements in this space and this is the first where I could not tell you, in a sentence, who owns the number that reaches the client statement. We discuss about it at every steering call and it never lands.”*
>
> **PORTFOLIO DIRECTOR**
>
> BPO, Payments, cards and disputes

Part 2 (Co-Management, section 1.4) reports the association of accountability with the four study measures.

2 Shared Risk and Issue Management

Shared risk and issue management were associated with the broader control environment. A single view of risk, one working issue process, timely resolution, and the ability to stop an agent address different parts of the same operational response. The findings concern how respondents assessed those functions, not simply whether each firm had a documented process. Their association supports examining coordination across the boundary between firms.

> *Review capacity has improved, I will say that. We have the hands now to check what the agents throw out each morning. What has not moved is stopping. If an agent starts matching breaks wrongly nobody can tell me who on either side is permitted to switch it off… we would still be mailing at noon."*
>
> **ENGAGEMENT MANAGER**
>
> BPO, Trade, settlement and reconciliation

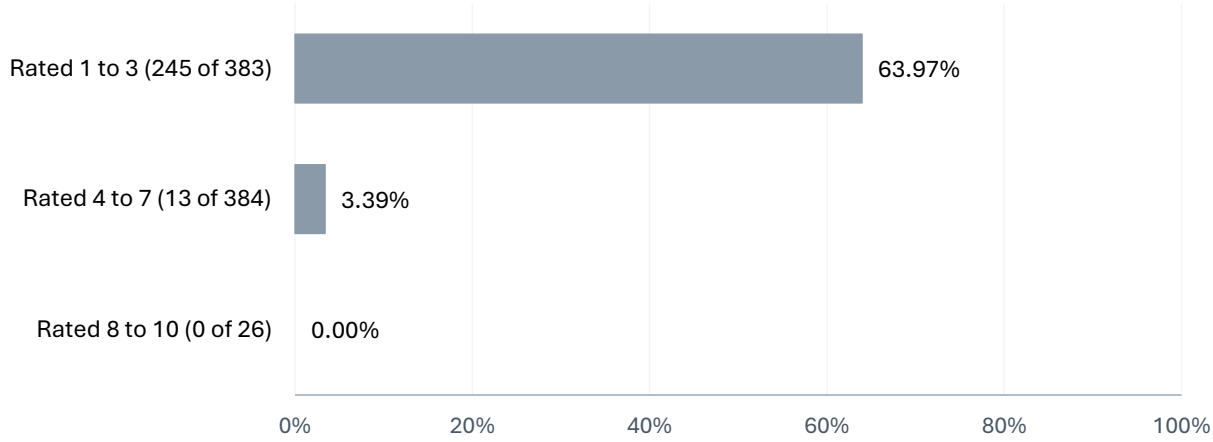
2.1 A Single Shared View of Agent Risk

Indicator C140 · Rate how well the two firms maintain a single shared view of the risks that agent work introduces · Full statistical analysis ›

A weak shared view of agent risk was associated with low scores. Separate risk records can leave the firms evaluating different parts of the same exposure, even where each maintains a formal risk process. The indicator concerns whether those views are brought together for joint oversight. The written accounts provide limited illustration of that arrangement, while the principal evidence comes from the ratings and their association with the broader score.

Figure 3.29. A Single Shared View of Agent Risk: share with a score of 3 or below by indicator rating.

A weak shared view of agent risk was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 383 respondents who rated the shared view of agent risk 1 to 3, 245 (63.97%) had a score of 3 or below. Among the 384 who rated it 4 to 7, 13 (3.39%) had a score of 3 or below, and none of the 26 who rated it 8 to 10 did (chi-square = 333.60, $p < 0.001$, Cramér's V = 0.65).

The shared view of risk was associated with the score at $r = 0.92$ (95% CI 0.89 to 0.94, n = 150 engagements) and with governance in the other four categories at $r = 0.78$. After adjustment for the mean of the other 53 indicators, the association remained at partial $r = 0.25$ ($p = 0.002$).

The association ranged across the active-stage groups, from 0.79 in Build-Out to 0.86 in Transition/Cut-Over and 0.89 in Production, and held within the canceled engagements at $r = 0.75$. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.87 among individual contributors, 0.89 among engagement managers, and 0.83 among portfolio directors. Among individual contributors alone, 64.76% (204 of 315) of

those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.71 (n = 383), against 5.24 (n = 410) among the rest (cluster-robust $p < 0.001$).

The written base was thin. Twenty-six respondents on 24 engagements wrote about a shared view of agent risk, 19 of the answers were negative, and 4 of the 19 described one of the 26 forms of a control bypassed or not working. The answers described two risk logs that met once a month, an exposure that remained on one side for a full cycle, and the same issue raised in two meetings and resolved in neither.

> *“my 1s here look harsh, mostly it is one thing… nobody keep a single risk list, so same issue get raise in two different meetings and fixed in neither. the people are polite, the process just does not exist”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Payments, cards and disputes

Part 2 (Co-Management, section 2.1) reports the association of the shared view of risk with the four study measures.

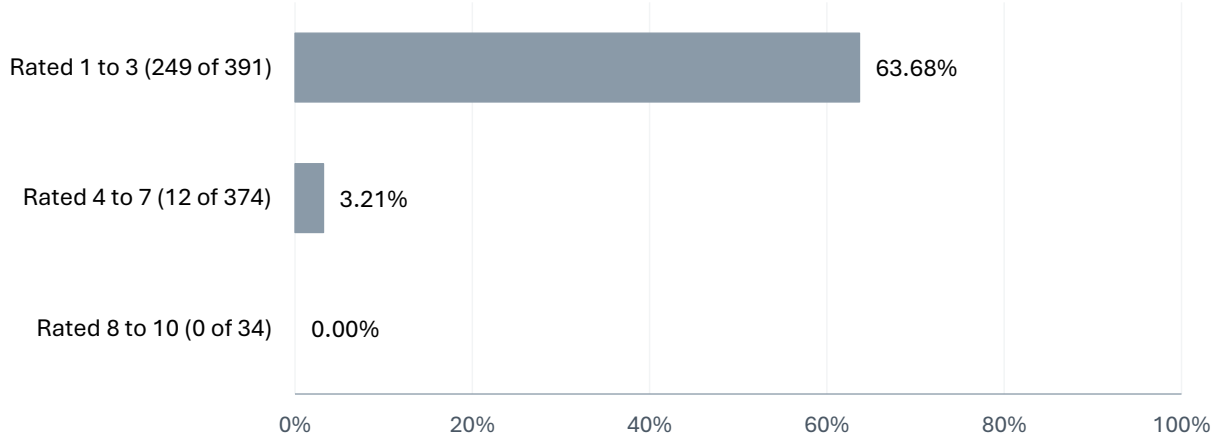
2.2 Co-Management of Issues Arising From Agent Work

Indicator C141 · Rate how effectively the two firms manage issues arising from agent work as one team rather than through two separate processes · Full statistical analysis ›

Fragmented issue management was associated with a weaker reported control environment. The relevant distinction is whether the two firms manage an agent-related problem together or pass it between separate processes. Shared management affects who can see the issue, assign action, and confirm resolution. The reported associations appear within both active and canceled groups, without establishing a change in this relationship over an engagement's life.

Figure 3.30. Co-Management of Issues Arising From Agent Work: share with a score of 3 or below by indicator rating.

Fragmented issue management was associated with a weaker reported control environment.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Within canceled engagements, r was 0.80. Only the visibility of each firm's agent actions to the other stood higher, at 0.82. Across the active stages the association ran at 0.82 in Build-Out and Transition/Cut-Over and 0.89 in Production, and after Holm adjustment it held in 85 of the 87 segment levels tested.

Scores were 3 or below for 249 of 391 (63.68%) at indicator ratings 1–3. The corresponding counts were 12 of 374 (3.21%) at indicator ratings 4–7 and 0 of 34 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 335.06, $p < 0.001$, Cramér's V = 0.65. The indicator was associated with the score at r = 0.92 (95% CI 0.89 to 0.94, n = 150 engagements) and with governance in the other four categories at r = 0.78. After adjustment for the mean of the other 53 indicators, the association remained at partial r = 0.26 (p = 0.001), the second-strongest adjusted association of the 20.

Within role the association stood at 0.87 among individual contributors and at 0.88 among both engagement managers and portfolio directors. Among individual contributors alone, 64.58% (206 of 319) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.71 (n = 391), against 5.30 (n = 408) among the rest (cluster-robust p < 0.001).

Fifty-four respondents on 50 engagements wrote about joint issue management, one answer each, and 39 of the answers were negative, 16 from canceled and 23 from active engagements. Of the 39, 11 (28.21%) described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (5). Respondents described two issue logs that agreed on perhaps half of their entries, merged queues agreed and never built, and alerts closed manually that appeared in no count. An engagement manager described two issue logs in which neither firm could see the other's agent actions, so each dispute became an argument over whose log was right.

> *“Two issue logs, theirs and ours. Neither side can see what the other's agents did. Every dispute becomes a discussion about whose record is right.”*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Technology engineering and operations

> *“we keep our own issue tracker and the bank compliance side keep theirs. When i log a false positive pattern it stays on our side, i have never once seen one cross over. And nobody has ever asked how many alerts we dispose by hand in a day, that number is written nowhere.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Financial crime and client onboarding

Part 2 (Co-Management, section 2.2) reports the association of joint issue management with the four study measures.

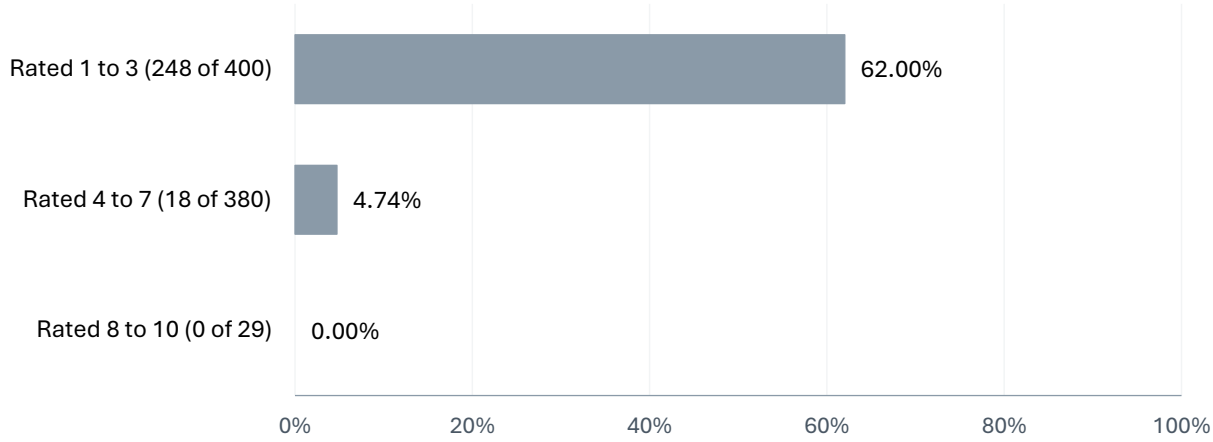
2.3 Resolution Speed Against the Spread of Agent-Caused Problems

Indicator C142 · Rate how well the speed of joint issue resolution matches the speed at which agent-caused problems can spread · Full statistical analysis ›

Low ratings of joint resolution speed were associated with low scores. For governance, the important comparison is between the time required to resolve an issue and the speed at which an agent-related problem can spread. The study measures respondents' assessment of that fit rather than elapsed incident-resolution times. The result supports examining operational response capacity while leaving the sequence and consequences of individual incidents unmeasured.

Figure 3.31. Resolution Speed Against the Spread of Agent-Caused Problems: share with a score of 3 or below by indicator rating.

Low ratings of joint resolution speed were associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Among respondents rating resolution speed at 4–7, 18 of 380 (4.74%) had a score of 3 or below. This was the second-highest share among the 20 co-management indicators, after human review capacity at 5.30%. Of the 400 respondents who rated resolution speed 1 to 3, 248 (62.00%) had a score of 3 or below, and none of the 29 who rated it 8 to 10 did (chi-square = 304.28, $p < 0.001$, Cramér's V = 0.61).

Resolution speed was associated with the score at $r = 0.91$ (95% CI 0.88 to 0.94, n = 150 engagements) and with governance in the other four categories at $r = 0.77$. After adjustment for the mean of the other 53 indicators, the association remained at partial $r = 0.25$ ($p = 0.002$).

The association was observed within the canceled engagements at $r = 0.78$ and across the active stages at 0.79 to 0.82, and after Holm adjustment it held in 86 of the 87 segment levels

tested. Within role it stood at 0.86 among both individual contributors and engagement managers and at 0.79 among portfolio directors. Among individual contributors alone, 63.38% (206 of 325) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.73 (n = 400), against 5.26 (n = 409) among the rest (cluster-robust $p < 0.001$).

The written base was thin. Eighteen respondents on 17 engagements wrote about resolution speed, and 16 of the answers were negative, 7 from canceled and 9 from active engagements. Three of the 16 described problems raised with no response. On one canceled engagement, 11 days passed before the two firms established whose change board owned an error, while three monthly attribution packs went to the financial institution's investment committee. On an active engagement the joint review met weekly, so a bad run was four days old before it was discussed, and the team kept a private list of the agent's changes in its place.

> *“In February the agents began handling the interaction effect differently after a platform update on our side — nobody had asked for that. It took eleven days to establish whose change board owned the question, by which time three monthly attribution packs had gone to the client's investment committee. Neither firm could show the other what had actually been done in that window. We closed it by recalculating the quarter by hand.”*
>
> **PORTFOLIO DIRECTOR**
>
> KPO, Trade, settlement and reconciliation

> *“the joint review is once a week and by then a bad run is four days old, so we run a private list of what the agent changed and revert back on chat when it looks off. that sheet is the real control, not the meeting”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

Part 2 (Co-Management, section 2.3) reports the association of resolution speed with the four study measures.

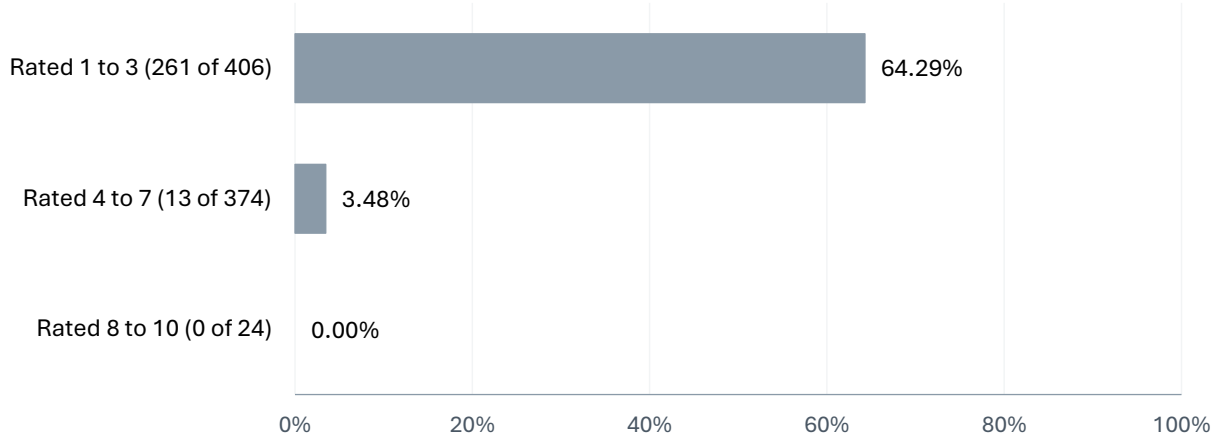
2.4 Speed of Stopping a Misbehaving Agent in Live Operations

Indicator C143 · Rate how quickly the two firms together can stop an agent that is behaving badly in live operations · Full statistical analysis ›

Limited ability to stop a misbehaving agent was associated with low scores, including a positive cross-category relationship under the adjusted model. This indicator concerns immediate operational control: the firms must be able to interrupt agent activity when a problem is recognized. It is distinct from routine approval, decision speed, and later remediation. The association does not estimate the causal benefit of a particular stop mechanism.

Figure 3.32. Speed of Stopping a Misbehaving Agent in Live Operations: share with a score of 3 or below by indicator rating.

Limited ability to stop a misbehaving agent was associated with low scores, including a positive cross-category relationship under the adjusted model.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

The adjusted cross-category estimate was partial r = 0.26 (nominal p = 0.001). The speed of stopping was associated with governance in the other four categories at r = 0.79, and its adjusted association (p = 0.001) was the third-strongest of the 20. It was associated with the score at r = 0.92 (95% CI 0.89 to 0.94, n = 150 engagements).

Scores were 3 or below for 261 of 406 (64.29%) at indicator ratings 1–3. The corresponding counts were 13 of 374 (3.48%) at indicator ratings 4–7 and 0 of 24 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 333.22, p < 0.001, Cramér's V = 0.64. Respondents who rated it distressed recorded a score of 2.70 (n = 406), against 5.28 (n = 398) among the rest (cluster-robust p < 0.001).

The association was observed within the canceled engagements at r = 0.78 and across the active stages at 0.81 to 0.87, and after Holm adjustment it held in 85 of the 87 segment levels tested. Within role it stood at 0.87 among individual contributors, 0.88 among engagement managers, and 0.82 among portfolio directors. Among individual contributors alone, 66.07% (220 of 333) of those who rated it 1 to 3 had a score of 3 or below.

The speed of stopping drew 73 written answers from 71 respondents on 51 engagements, and 59 were negative, 21 from canceled and 38 from active engagements. Of the 59, 13 (22.03%) described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (8) and the absence of an escalation path outside working hours (4). Respondents described three different names given for the person who could switch an agent off, and overnight shifts without a person authorized to pause one.

> *“the thing the numbers don't show is nobody could actually switch the thing off. i asked who does that and got three different names”*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Corporate and back-office functions

> *“problem is at night no one over there can authorise stopping the triage bot, i have to raise ticket and wait for morning shift. one named person each side per shift with authority to pause the same, that is all it needs. the alerts it auto closes are the risk, not the ones it leaves open.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Technology engineering and operations

Part 2 (Co-Management, section 2.4) reports the association of the speed of stopping with the four study measures.

“

In march we change the tone rule for one client segment. i put in a change request. It sit for five week. In that time the tool send out maybe forty letters on the old rule and nobody flag it. When the request finally come back approved, the rule is already changed by somebody else, a different way. Nobody tell me. i find out from another analyst. There is no one risk list, both side keep their own”

INDIVIDUAL CONTRIBUTOR

CXM, Client servicing and communications

3 Joint Change Control

The three change-control indicators connect the pace and visibility of agent change with the wider score. They distinguish a joint process that can keep pace, notice that reaches the other firm before effect, and governance of behavior between formal releases. These are complementary control functions. A strong association for one does not establish that the others can be treated as substitutes or that their relative coefficients rank causal importance.

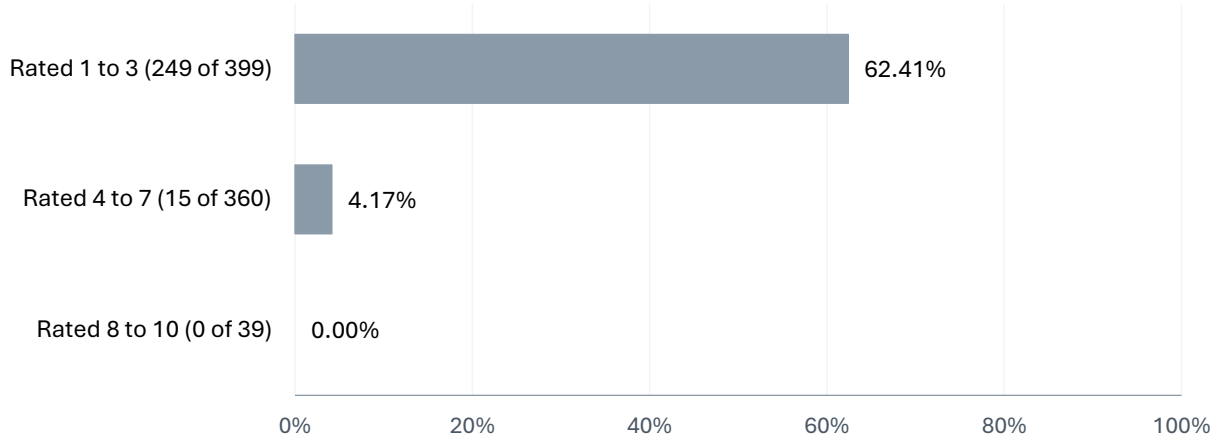
> *"What we do now is we keep a group chat and that is the real record. Anything that needs a decision goes there, somebody screenshots it, and that becomes the approval. Nobody signed anything. If you ask me who owns the rule change from last month I will show you a chat message from a Tuesday, that is all there is. On paper there is a joint forum every two weeks. Half the time it is cancelled, the other half they send somebody who takes notes and says he will come back to us. He does not come back. So we made our own list, three of us keep it, we tick items off and if nobody objects in two days we go ahead. That is not governance, I know that, but the work has to move and no one else will move it. The other thing is the review. It is me reviewing my teammate and him reviewing mine, we are both tired and we both want to go home, so the second look is not really a second look. I raise this, feedbacks go up, nothing come down. Then when something breaks the question is always why did nobody check, and I do not have a good answer for that anymore."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Regulatory, audit, tax and legal

3.1 Joint Change Control Against the Speed of Agent Behavior Change

Indicator C144 · Rate how well the joint change control process keeps pace with the speed of agent behavior changes · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 3.8 · Full statistical analysis ›

Joint change-control ratings retained a positive cross-category association under the specified adjustment. The operational concern is whether the two firms can govern changes at the pace at which agent behavior changes, rather than relying on a formal process whose cycle is too slow for the work. The observed coefficient is among the largest positive estimates in this comparison. Its ordering does not establish that this control has the largest causal effect.

Figure 3.33. Joint Change Control Against the Speed of Agent Behavior Change: share with a score of 3 or below by indicator rating.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Change control was associated with governance in the other four categories at r = 0.79, and after that adjustment the association remained at partial r = 0.31 ($p < 0.001$). It was associated with the score at r = 0.92 (95% CI 0.89 to 0.94, n = 150 engagements), and after Holm adjustment the association held in 87 of the 87 segment levels tested.

Scores were 3 or below for 249 of 399 (62.41%) at indicator ratings 1–3. The corresponding counts were 15 of 360 (4.17%) at indicator ratings 4–7 and 0 of 39 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 310.22, $p < 0.001$, Cramér's V = 0.62. Its 1 to 3 band share was among the lowest of the 20 co-management indicators, while its adjusted association was the highest. These comparisons answer different questions. The rating bands describe the distribution of low governance, oversight, and control scores. The adjusted

association describes the relationship remaining after accounting for the comparison mean. Their rankings do not provide a common scale of practical importance.

The association was observed within the canceled engagements at r = 0.79, the third-strongest of the 20, and ranged across the active-stage groups from 0.79 in Build-Out to 0.87 in Production. Within role it stood at 0.87 among individual contributors, 0.89 among engagement managers, and 0.78 among portfolio directors. Among individual contributors alone, 63.41% (208 of 328) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.74 (n = 399), against 5.25 (n = 399) among the rest (cluster-robust $p < 0.001$).

Joint change control drew 71 written answers from 70 respondents on 53 engagements, and 57 were negative, 25 from canceled and 32 from active engagements. Of the 57, 14 (24.56%) described one of the 26 forms of a control bypassed or not working, most often change discovered only through failure (8) and problems raised with no response (5). Respondents described a change request that waited five weeks while the agent kept working on the old rule, and changes that waited for a joint decision and then had to be made in haste.

> *“they never decide anything together. a mid-term change sit two weeks and then we are told to fix it fast”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Client servicing and communications

Part 2 (Co-Management, section 3.1) reports the association of joint change control with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 3.8) its association with manual alteration.

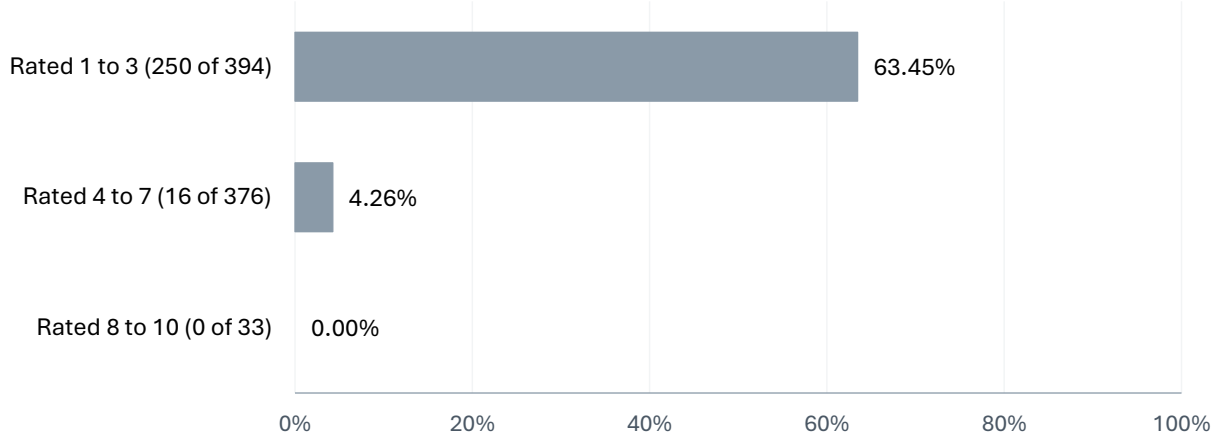
3.2 Visibility of Agent Updates to the Other Firm Before They Take Effect

Indicator C145 · Rate how well agent updates made by one firm become visible to the other firm before they take effect · Full statistical analysis ›

Poor advance visibility of agent updates was associated with low scores. The control concerns whether one firm's change becomes known to the other before it affects shared work. Notification after a change takes effect serves a different purpose from an opportunity to assess and prepare for it. The findings connect reported update visibility with the wider control environment without directly measuring every update or response.

Figure 3.34. Visibility of Agent Updates to the Other Firm Before They Take Effect: share with a score of 3 or below by indicator rating.

Poor advance visibility of agent updates was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 394 respondents who rated the visibility of agent updates 1 to 3, 250 (63.45%) had a score of 3 or below. Among the 376 who rated it 4 to 7, 16 (4.26%) had a score of 3 or below, and none of the 33 who rated it 8 to 10 did (chi-square = 321.39, $p < 0.001$, Cramér's V = 0.63).

The visibility of updates was associated with the score at $r = 0.91$ (95% CI 0.88 to 0.94, n = 150 engagements) and with governance in the other four categories at $r = 0.77$. After adjustment for the mean of the other 53 indicators, the association remained at partial $r = 0.25$ ($p = 0.002$).

The association was observed within the canceled engagements at $r = 0.78$, the fifth-strongest of the 20, and across the active stages at 0.82 to 0.84. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.87 among individual contributors, 0.88 among engagement managers, and 0.80 among portfolio directors. Among

individual contributors alone, 64.51% (209 of 324) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.72 (n = 394), against 5.23 (n = 409) among the rest (cluster-robust $p < 0.001$).

Forty-eight respondents on 44 engagements wrote about the visibility of updates, one answer each, and 38 of the answers were negative, 12 from canceled and 26 from active engagements. Of the 38, 17 (44.74%) described one of the 26 forms of a control bypassed or not working, the highest proportion among the co-management indicators with 30 or more answers, and 14 of the 17 described change discovered only through failure. Respondents learned of the other firm's updates from failed cases and from figures that moved. A portfolio director reported that decision rights had been written down while updates still went live before the financial institution's conduct team knew of them.

> *"one point only. when they change something on the agent side we find out from the output, not from anybody telling us"*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Payments, cards and disputes

> *"The joint forum has come a long way since the spring - decision rights on which alerts the agents may close without a reviewer are now written down and both sides work to the same list, and that argument used to take two weeks each time it came up. What has not moved is visibility of changes. Twice this quarter a tuning change on our side went live before the bank's conduct team knew of it, and while neither was harmful, we would have no fast way to pull an agent back mid-shift if one ever were. Kindly note that is a build neither side has funded."*
>
> **PORTFOLIO DIRECTOR**
> CXM, Regulatory, audit, tax and legal

Part 2 (Co-Management, section 3.2) reports the association of the visibility of updates with the four study measures.

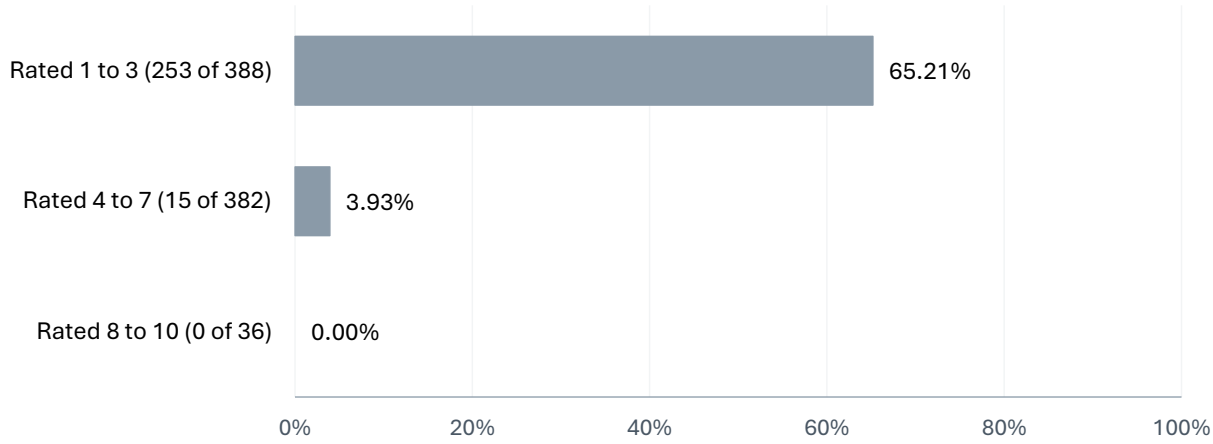
3.3 Joint Governance of Agent Behavior Change Between Formal Releases

Indicator C146 · Rate how well the two firms jointly govern changes agents make to their own behavior between formal releases · Full statistical analysis ›

Weak joint governance of behavior changes between releases was associated with low scores. This extends the change-control question beyond formally scheduled updates to behavior that changes while the system remains in operation. A release process can be well defined while this intervening responsibility remains unclear. The indicator adds a distinct control question even though its statistical pattern resembles those of other co-management indicators.

Figure 3.35. Joint Governance of Agent Behavior Change Between Formal Releases: share with a score of 3 or below by indicator rating.

Weak joint governance of behavior changes between releases was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 388 respondents who rated it 1 to 3, 253 had a score of 3 or below, the same count and share as for step ownership. Among the 382 who rated it 4 to 7, 15 (3.93%) had a score of 3 or below, and none of the 36 who rated it 8 to 10 did (chi-square = 344.45, $p < 0.001$, Cramér's V = 0.65).

The governance of change between releases was associated with the score at r = 0.90 (95% CI 0.87 to 0.93, n = 150 engagements), the joint weakest of the 20, and with governance in the other four categories at r = 0.76. After adjustment for the mean of the other 53 indicators, the association remained at partial r = 0.23 (p = 0.004).

The association was observed within the canceled engagements at r = 0.75 and across the active stages at 0.80 to 0.84, and after Holm adjustment it held in 86 of the 87 segment levels

tested. Within role it stood at 0.86 among individual contributors and portfolio directors and at 0.85 among engagement managers. Among individual contributors alone, 67.20% (211 of 314) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.70 (n = 388), against 5.22 (n = 418) among the rest (cluster-robust $p < 0.001$).

Fourteen respondents on 14 engagements wrote about change between releases, and 12 of the answers were negative, 4 from canceled and 8 from active engagements. Three of the 12 described one of the 26 forms of a control bypassed or not working. Respondents described agent behavior that drifted between releases with no forum quick enough to rule on whether a drift was a fix or a defect, so that the delivery team decided and informed both firms afterward. An individual contributor reported change control faster than on earlier work and no agreed definition of a defect when a configuration retuned itself.

> *"The scale does not really capture timing. On paper the joint model works — shared risk log, weekly review, both sides turn up prepared, and the quality checks do catch the errors the agents actually make. What it cannot show is that agent behavior drifts a little between releases and there is no forum that meets fast enough to rule on whether a drift is a fix or a defect. We end up deciding on the desk and telling both sides afterward."*
>
> **ENGAGEMENT MANAGER**
>
> CXM, Client servicing and communications

> *"Both sides turn up when a Feed goes bad, the reviewing keeps pace with what the agents produce, moreover change control is faster here than what i saw before. Where it slips is we dont have an agreed definition of a defect when the self healing config retunes itself, so it sits in two trackers and each side thinks the other one has it"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

Part 2 (Co-Management, section 3.3) reports the association of the governance of change between releases with the four study measures.

“

One issue dominates and I will leave the others alone. The human review point on arrangement offers has quietly become an approval step rather than a scrutiny step. The reviewer opens a queue of forty and releases thirty-eight, which is not review in any sense a supervisor would accept. That is a capacity design problem and not a people problem. The governance we jointly operate was drawn up for a world in which a person made each offer and could be asked afterwards why, and it does not sit comfortably over work of this shape, however well both sides chair the meetings.”

PORTFOLIO DIRECTOR

CXM, Payments, cards and disputes

4 Human Review and the Joint Quality Assurance Loop

Human review was associated with the wider score across placement, capacity, scrutiny, error capture, and measurement. These indicators distinguish a formal review point from a review system that can examine the relevant volume of work and identify consequential errors. Their common pattern supports examining the operation of the whole review arrangement. Each indicator remains necessary because a weakness in one dimension is not described by the others alone.

> *“the problem is review is only a signature now. risk register did get better, we go through it together every fortnight and both sides see same list. but on code review two people tick it in ten minutes and nobody checks what the agent actually changed”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

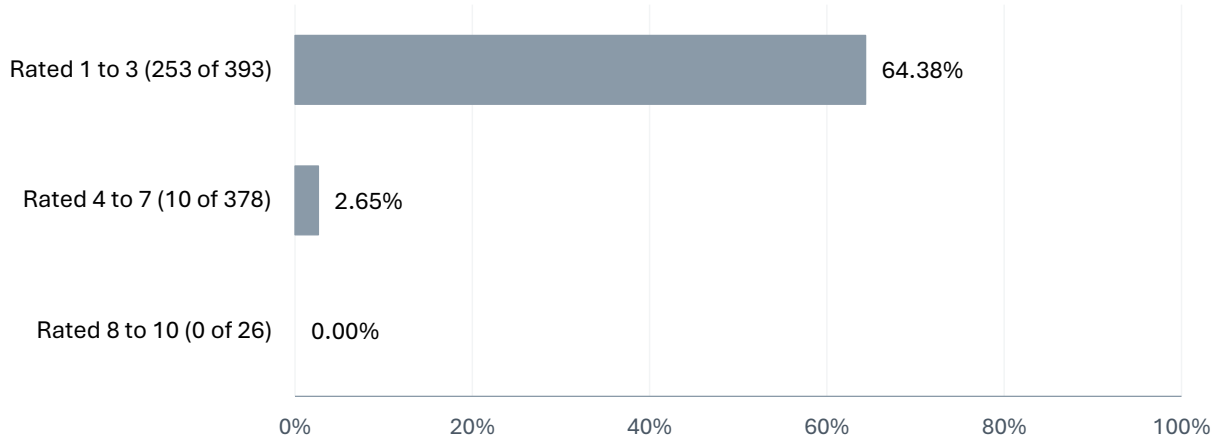
4.1 Placement of Human Review Points

Indicator C147 · Rate how well human review points are placed across the two firms at the moments where agent errors would matter most · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 6.2 · Full statistical analysis ›

Poor placement of human review points was associated with low scores. Review contributes to control only if it occurs where an error can still be identified and addressed before its consequences matter. This indicator concerns that placement, rather than the number of reviewers or the scrutiny they apply. Its rating-band comparisons describe the reported relationship and do not validate a boundary between working and failed governance.

Figure 3.36. Placement of Human Review Points: share with a score of 3 or below by indicator rating.

Poor placement of human review points was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Scores were 3 or below for 253 of 393 (64.38%) at indicator ratings 1–3. The corresponding counts were 10 of 378 (2.65%) at indicator ratings 4–7 and 0 of 26 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 345.33, $p < 0.001$, Cramér's V = 0.66.

Placement was associated with the score at $r = 0.91$ (95% CI 0.88 to 0.94, n = 150 engagements) and with governance in the other four categories at $r = 0.77$. After adjustment for the mean of the other 53 indicators, the association remained at partial $r = 0.23$ ($p = 0.004$).

The association was observed within the canceled engagements at $r = 0.78$ and ranged across the active-stage groups from 0.77 in Build-Out to 0.85 in Production. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.86 among individual contributors and engagement managers and at 0.88 among portfolio directors.

Among individual contributors alone, 65.83% (210 of 319) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.72 (n = 393), against 5.28 (n = 404) among the rest (cluster-robust $p < 0.001$).

Placement tracked the manual alteration of AI outputs across its rating bands. Of the respondents who rated placement 1 to 3, 52.65% (199 of 378) altered AI outputs, against 13.98% (52 of 372) at ratings of 4 to 7 and 3 of 25 at ratings of 8 to 10 (chi-square $p < 0.001$). Part 5 (Leading Dynamics of Manual Alteration, section 6.2) holds the crossing of placement with manual alteration in full.

Placement drew 105 written answers from 103 respondents on 75 engagements, and 75 were negative, 30 from canceled and 45 from active engagements. Of the 75, 14 (18.67%) described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (9) and change discovered only through failure (4). Respondents described review that sat after validation or after the output had left, tickets that closed with neither firm checking them, and an agent that rewrote consent flags on nine thousand records overnight without approval.

> *"Neither side checks a ticket before it closes. Thats all."*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Corporate and back-office functions

> *"In March the agent rewrites consent flags on nine thousand records overnight. You are asking me about accuracy and memory, but the thing that mattered was it acted at all, without anyone aproving it."*
>
> **ENGAGEMENT MANAGER**
> ITO, Technology engineering and operations

Part 2 (Co-Management, section 4.1) reports the association of placement with the four study measures.

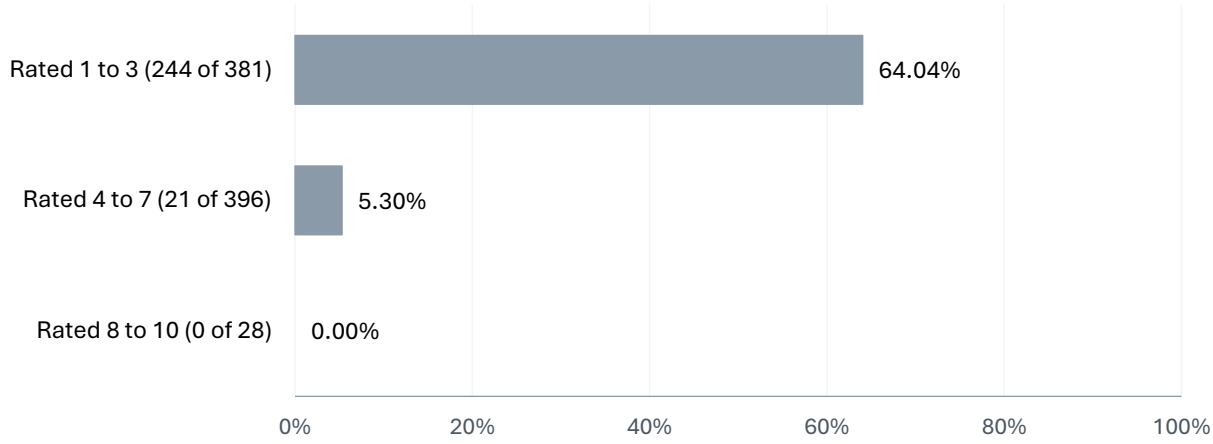
4.2 Human Review Capacity Against Agent Output Volume

Indicator C148 · Rate how well joint human review capacity keeps up with the volume of output the agents produce · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 6.3 · Full statistical analysis ›

Insufficient human review capacity was associated with low scores. The control issue is whether the volume of agent output can be reviewed by the people assigned to oversee it. Capacity differs from review placement and from genuine scrutiny: a correctly placed review point may still be overwhelmed. The written accounts and ratings address that operational mismatch, while the score association does not identify a sufficient staffing level.

Figure 3.37. Human Review Capacity Against Agent Output Volume: share with a score of 3 or below by indicator rating.

Insufficient human review capacity was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 396 respondents who rated review capacity 4 to 7, 21 (5.30%) had a score of 3 or below. Among the 381 who rated it 1 to 3, 244 (64.04%) had a score of 3 or below, and none of the 28 who rated it 8 to 10 did (chi-square = 317.63, $p < 0.001$, Cramér's V = 0.63).

Review capacity was associated with the score at r = 0.90 (95% CI 0.86 to 0.92, n = 150 engagements), the joint weakest of the 20, and with governance in the other four categories at r = 0.75, the weakest of the 20. After adjustment for the mean of the other 53 indicators, the association remained at partial r = 0.25 (p = 0.003).

The association was observed within the canceled engagements at r = 0.76 and ranged across the active-stage groups from 0.75 in Build-Out to 0.82 in Transition/Cut-Over and Production. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it

stood at 0.86 among individual contributors and 0.83 among engagement managers and portfolio directors. Among individual contributors alone, 66.03% (206 of 312) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.71 (n = 381), against 5.18 (n = 424) among the rest (cluster-robust $p < 0.001$).

Review capacity tracked human rework. Of the respondents who rated capacity 1 to 3, 43.57% (166 of 381) described human rework absorbing agent failure, against 14.90% (59 of 396) at ratings of 4 to 7 and 1 of 28 at ratings of 8 to 10 (chi-square $p < 0.001$). The written pattern of review overrun by agent volume is analyzed with the other reasons for reported governance weakness under The Fit Between Governance Processes and the Pace of Agent Work.

Review capacity drew 145 written answers from 135 respondents on 89 engagements, and 134 of them were negative, 42 from canceled and 92 from active engagements. Of the 134, 17 (12.69%) described one of the 26 forms of a control bypassed or not working, most often unclear liability or regulatory accountability (6). Respondents described review that grew shorter as volume grew, a formal quality sample of about two percent, and one or two people as the whole of review across several workstreams.

> *“I edit the suggested response before the Agent sends it, it still guesses wrong on routing a lot. And QA on this is basically me reading things. There is a formal QA step, it samples about two percent, which means ninety eight percent of what goes out has had exactly one set of eyes on it and those eyes are mine and I am also handling my own queue. That is the honest picture. Nobody is hiding it, its just that nobody has added it up. The agent handles the easy ones well, genuinely, and for the routine balance and payment questions its better than we were. Its the edge cases where it confidently gets it wrong, and confidently is the part that matters, because it does not flag uncertainty, it just answers. So you have to read all of them to find the few. And the productivity numbers we report do not account for the reading. I have brought this up with my supervisor twice. Both times the response was that the sampling rate would be reviewed. I do not think anyone is being”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Client servicing and communications

> *“one reviewer for forty provisioning outputs in a day, she cant”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

Part 2 (Co-Management, section 4.2) reports the association of review capacity with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 6.3) its association with manual alteration.

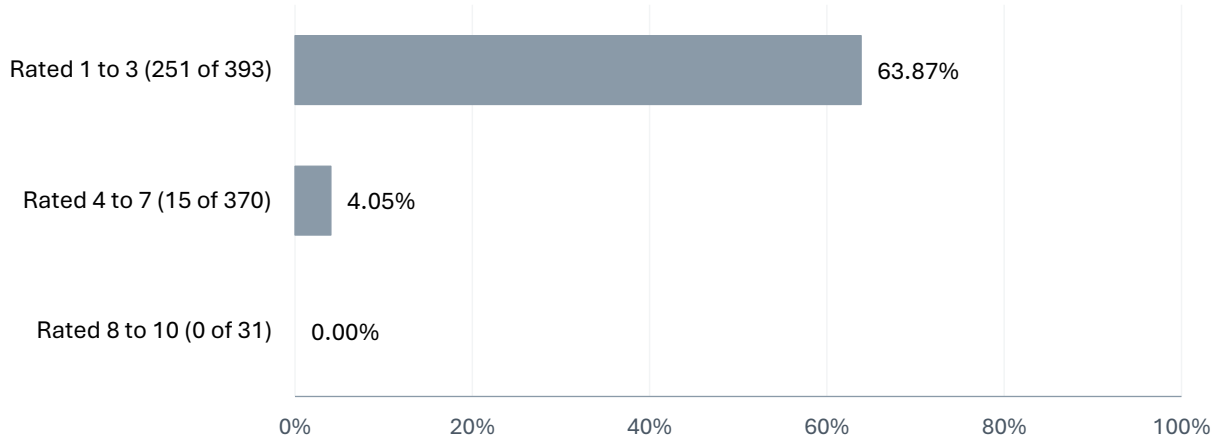
4.3 Genuine Scrutiny at Human Review Points

Indicator C149 · Rate how much genuine scrutiny agent outputs receive at human review points rather than routine approval · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 6.4 · Full statistical analysis ›

Routine approval without substantial scrutiny was associated with a weaker reported control environment. The indicator asks whether human review operates as an examination of agent work or primarily as a procedural sign-off. It gives the governance analysis a quality dimension that review placement and capacity do not capture alone. Differences among subgroup correlations remain descriptive and do not establish that scrutiny deteriorated on cancellation.

Figure 3.38. Genuine Scrutiny at Human Review Points: share with a score of 3 or below by indicator rating.

Routine approval without substantial scrutiny was associated with a weaker reported control environment.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Within canceled engagements, the correlation was 0.74, jointly the lowest of the 20 with step ownership, while across the active stages it ran at 0.81 in Build-Out, 0.83 in Transition/Cut-Over, and 0.86 in Production.

Scores were 3 or below for 251 of 393 (63.87%) at indicator ratings 1–3. The corresponding counts were 15 of 370 (4.05%) at indicator ratings 4–7 and 0 of 31 (0.00%) at indicator ratings

8–10. The reported band results were chi-square = 322.30, p < 0.001, Cramér's V = 0.64. Scrutiny was associated with the score at r = 0.92 (95% CI 0.89 to 0.94, n = 150 engagements) and with governance in the other four categories at r = 0.78. After adjustment for the mean of the other 53 indicators, the association remained at partial r = 0.25 (p = 0.002).

After Holm adjustment the association held in 86 of the 87 segment levels tested. Within role it stood at 0.87 among individual contributors, 0.86 among engagement managers, and 0.83 among portfolio directors. Among individual contributors alone, 65.62% (208 of 317) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.71 (n = 393), against 5.27 (n = 401) among the rest (cluster-robust p < 0.001).

Genuine scrutiny drew 66 written answers from 63 respondents on 50 engagements, and 59 were negative, 30 from canceled and 29 from active engagements. Of the 59, 11 (18.64%) described one of the 26 forms of a control bypassed or not working, most often activity counted in place of outcome (4), change discovered only through failure (3), and problems raised with no response (3). Respondents described releases approved without reading, approval by the same person who had prepared the case, and a queue of forty offers from which thirty-eight were released.

> *“Nobody asked this one, the person who approve our HR cases is the same person who prepare them, so the review is only a click. Please note also when they change the routing rules nobody tell us, we find out from the failed cases”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Corporate and back-office functions

Part 2 (Co-Management, section 4.3) reports the association of scrutiny with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 6.4) its association with manual alteration.

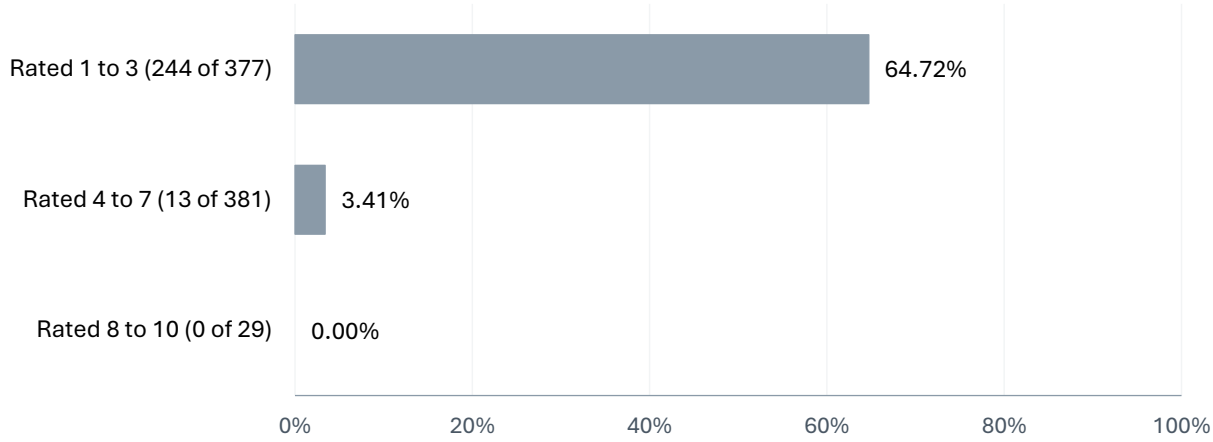
4.4 Reliability of the Joint Quality Assurance Loop

Indicator C150 · Rate how reliably the joint quality assurance loop catches agent errors before they reach live operations · Full statistical analysis ›

A less reliable joint quality assurance loop was associated with low scores. This concerns whether the combined process catches agent errors before they reach live operations, connecting review activity with its reported operational result. It differs from the existence of a review step or a formal quality policy. The adjusted association remains conditional on the comparison measure, with no inference that its relative ranking measures the control's practical importance.

Figure 3.39. Reliability of the Joint Quality Assurance Loop: share with a score of 3 or below by indicator rating.

A less reliable joint quality assurance loop was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

The quality loop was associated with governance in the other four categories at r = 0.77, and after adjustment for the mean of the other 53 indicators the association remained at partial r = 0.21 (p = 0.012). It was associated with the score at r = 0.91 (95% CI 0.88 to 0.94, n = 150 engagements).

The weak adjusted association sat beside one of the sharpest band crossings of the 20. Scores were 3 or below for 244 of 377 (64.72%) at indicator ratings 1–3. The corresponding counts were 13 of 381 (3.41%) at indicator ratings 4–7 and 0 of 29 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 338.49, p < 0.001, Cramér's V = 0.66.

The association was observed within the canceled engagements at r = 0.77 and ranged across the active-stage groups from 0.78 in Build-Out to 0.86 in Production. After Holm adjustment, it

was significant in 86 of the 87 segment levels tested. Within role it stood at 0.86 among individual contributors and engagement managers and at 0.76 among portfolio directors. Among individual contributors alone, 67.55% (204 of 302) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.71 (n = 377), against 5.22 (n = 410) among the rest (cluster-robust $p < 0.001$).

The written base was thin. Twenty-nine respondents on 28 engagements wrote about the quality loop, and 22 of the answers were negative, 11 from canceled and 11 from active engagements. Of the 22, 7 (31.82%) described one of the 26 forms of a control bypassed or not working, most often change discovered only through failure (3) and problems raised with no response (3). Respondents described errors found only once they were live, and drafts that went out wrong without anyone stopping them.

> *I give 1 for one reason, no one ever stop a draft going out wrong"*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Client servicing and communications

> *My 4 is only for the QA bit, the errors we find are allready live by then"*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Client servicing and communications

Part 2 (Co-Management, section 4.4) reports the association of the quality loop with the four study measures.

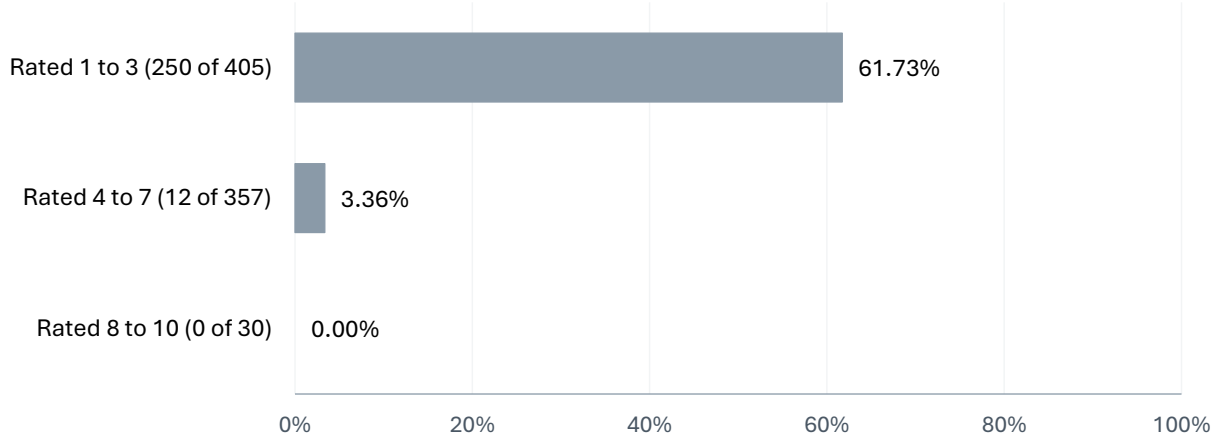
4.5 Quality Measures That Capture Agent Errors

Indicator C151 · Rate how well the quality measures both firms use capture the kinds of errors agents actually make · Full statistical analysis ›

Quality measures that inadequately captured agent errors were associated with low scores. The measures used by both firms need to capture the errors agents actually make. Counts of completed work or cleared exceptions can be informative while leaving incorrect completions poorly represented. The limited written accounts illustrate that distinction, without establishing an exhaustive inventory of errors or the independent accuracy of the measures.

Figure 3.40. Quality Measures That Capture Agent Errors: share with a score of 3 or below by indicator rating.

Quality measures that inadequately captured agent errors were associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

The indicator was associated with the score at r = 0.93 (95% CI 0.90 to 0.95, n = 150 engagements), third among the 47 governance indicators. It was associated with governance in the other four categories at r = 0.79, and after adjustment for the mean of the other 53 indicators the association remained at partial r = 0.26 (p = 0.002), the fifth-strongest of the 20.

Scores were 3 or below for 250 of 405 (61.73%) at indicator ratings 1–3. The corresponding counts were 12 of 357 (3.36%) at indicator ratings 4–7 and 0 of 30 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 307.41, p < 0.001, Cramér's V = 0.62. The 1 to 3 band share was the second-lowest of the 20, the pattern seen for accountability, where a close tie across the whole range sat beside low ratings spread more widely across engagements.

The association was observed within the canceled engagements at r = 0.77 and across the active stages at 0.81 to 0.85, and after Holm adjustment it held in 85 of the 87 segment levels tested. Within role it stood at 0.87 among individual contributors, 0.88 among engagement managers, and 0.86 among portfolio directors. Among individual contributors alone, 63.53% (209 of 329) of those who rated them 1 to 3 had a score of 3 or below. Respondents who rated them distressed recorded a score of 2.73 (n = 405), against 5.33 (n = 387) among the rest (cluster-robust p < 0.001).

Fourteen respondents on 14 engagements wrote about quality measures, and 11 of the answers were negative, 6 from canceled and 5 from active engagements. Five of the 11 described one of the 26 forms of a control bypassed or not working, most often change discovered only through failure (3). The answers placed the failure in measures designed for the errors people make. Those measures counted an exception cleared as a pass whether or not it had been cleared correctly. An individual contributor noticed an unannounced change to an audit sampling rule only when the exception count halved overnight.

> *“There was a sampling rule that changed on a continuous audit run and nobody on either side told us, i only noticed because exception count dropped by half overnight and that is not normal for that control. When i asked, neither side would say who approved it. Their audit lead said it is vendor change, my lead said it is their side, and i am still not knowing which is true. On my previous engagement, which was reconciliation work, this was different. Not perfect but there was one change log and both firms wrote into it, so if something moved you could at least see who moved it and when. Here there is nothing like the same. Changes just appear. And second thing is the qa checklist, it is written for a human tester. It asks did you check the sample, did you sign it, it does not ask did the tool behave the way it behaved last week, because nobody thought to put that question in. So we tick boxes that mean nothing and the real problem goes straight through. Kindly note i did raise this with my lead in Janurary. Nothing came back. Previous engagement was a small team and a decision took a day, here a decision needs three names and it takes a month or it never comes at all, and then we just decide ourself and carry on, which everybody knows and nobody writes down.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Regulatory, audit, tax and legal

Part 2 (Co-Management, section 4.5) reports the association of quality measures with the four study measures.

“

I never raise this with anyone above me, just sort duplicate myself and move on… And this happens on most files now, on top of that source system clearly has no dedup on their side, we are the dedup”

INDIVIDUAL CONTRIBUTOR

KPO, Credit, underwriting and suitability

5 The Shift From Managing People to Managing Agents

The fit between co-management and agent-performed work was associated with the wider score. Cross-firm visibility, decision speed, present operating-model fit, and readiness for expected autonomy describe different demands on that management arrangement. The evidence connects those reported demands with governance quality, while the future-autonomy item remains an expectation rather than observed future capability.

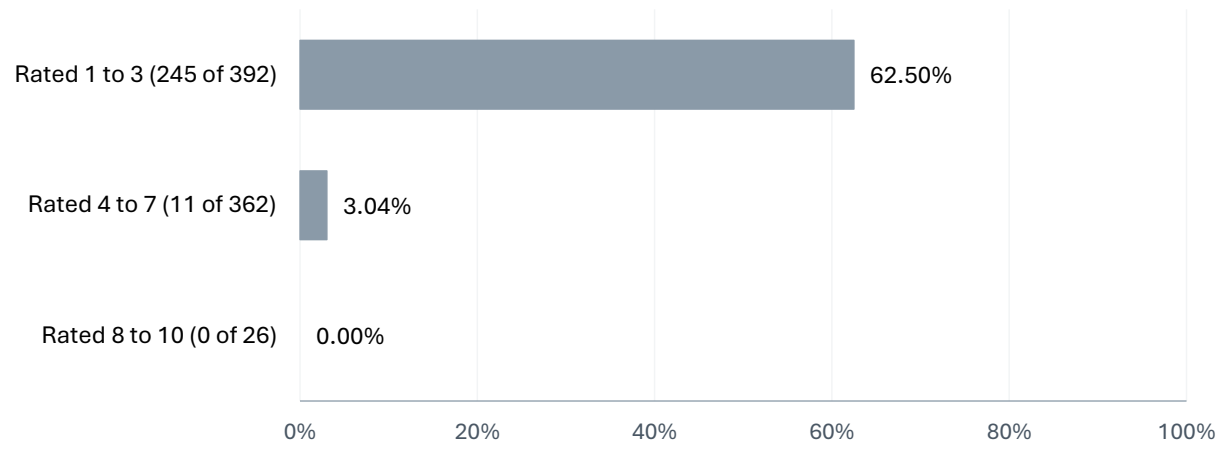
5.1 Each Firm's Visibility of the Other Firm's Agent Actions

Indicator C152 · Rate how fully each firm can see what the other firm's agents actually did when questions arise · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 6.7 · Full statistical analysis ›

Limited visibility of the other firm's agent actions was associated with low scores. Oversight requires an account of what occurred across the shared workflow when a question or incident arises. This indicator concerns cross-firm access to that account, beyond whether a tool generates logs internally. The association was observed within active and canceled groups, with no direct test of how visibility changed over time.

Figure 3.41. Each Firm's Visibility of the Other Firm's Agent Actions: share with a score of 3 or below by indicator rating.

Limited visibility of the other firm's agent actions was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

The association within the canceled engagements matched the association across the active stages, at 0.82 in Build-Out, 0.84 in Transition/Cut-Over, and 0.83 in Production, with no direct test of change over time. After Holm adjustment, it was significant in 86 of the 87 segment levels tested.

Scores were 3 or below for 245 of 392 (62.50%) at indicator ratings 1–3. The corresponding counts were 11 of 362 (3.04%) at indicator ratings 4–7 and 0 of 26 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 314.93, $p < 0.001$, Cramér's V = 0.64. Visibility was associated with the score at $r = 0.92$ (95% CI 0.89 to 0.94, n = 150 engagements) and with governance in the other four categories at $r = 0.79$. After adjustment for the mean of the other 53 indicators, the association remained at partial $r = 0.22$ ($p = 0.007$).

Within role the association stood at 0.88 among individual contributors, 0.87 among engagement managers, and 0.85 among portfolio directors. Among individual contributors alone, 64.86% (203 of 313) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.74 (n = 392), against 5.35 (n = 388) among the rest (cluster-robust $p < 0.001$).

Visibility drew 66 written answers from 64 respondents on 49 engagements, and 62 of them were negative, 23 from canceled and 39 from active engagements, the most uniformly negative testimony among the co-management indicators. Of the 62, 9 (14.52%) described one of the 26 forms of a control bypassed or not working, most often unclear liability or regulatory accountability (3) and exceptions omitted from the reports the financial institution received (2). Respondents described agent actions reconstructed from logs after the event and answers that were, by the time they were given, partly guessed.

> *“I will say this for the joint side, when we do get their AML lead in a room the decisions are quick and they hold, and their reviewers keep up with the volume we send, which I did not expect. The failure is not the people, it is that neither firm has sight of the other side's agent actions. They ask me on Tuesday why an alert closed on Sunday and I reconstruct it from logs, and by then the answer is half a guess. Nobody designed for that question.”*
>
> **ENGAGEMENT MANAGER**
>
> BPO, Financial crime and client onboarding

> *When something the agent did needs explaining i cant see there side of it, so i rebuild the break from the custodian statement myself and work back from that. Takes a morning. My team does the same. Nobody reviews the volume that comes out, we just spot check the big ones"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Trade, settlement and reconciliation

Part 2 (Co-Management, section 5.1) reports the association of visibility with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 6.7) its association with manual alteration.

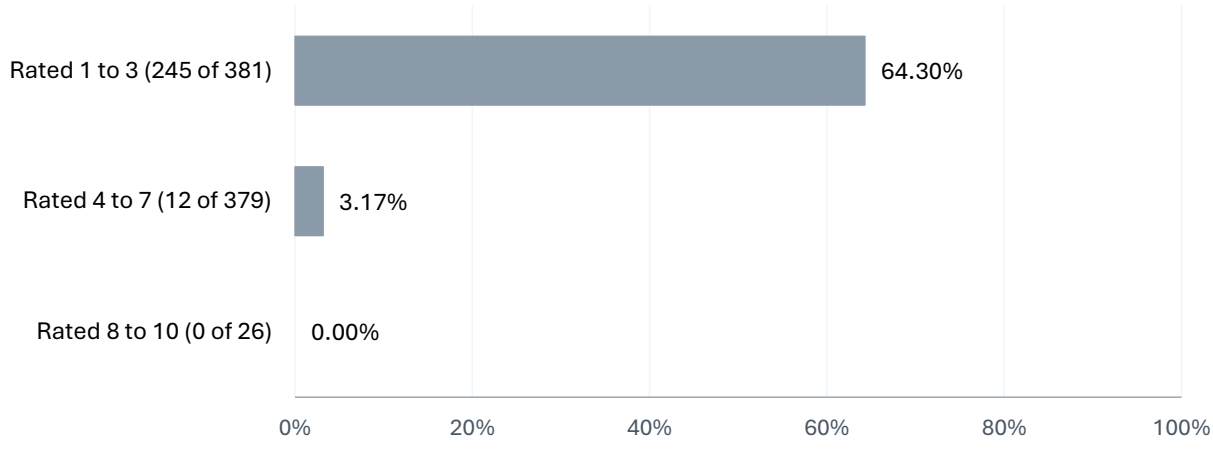
5.2 Joint Decision Speed Against Agent Speed

Indicator C153 · Rate how closely the speed of joint decision making matches the speed at which the agents act · Full statistical analysis ›

Joint decisions that did not keep pace with agent action were associated with low scores. The governance issue is the time available to decide when operational work continues faster than the two firms can reach agreement. This differs from deciding who has authority in principle. The evidence describes reported decision speed and its relationship with the broader score, not a measured delay for every decision.

Figure 3.42. Joint Decision Speed Against Agent Speed: share with a score of 3 or below by indicator rating.

Joint decisions that did not keep pace with agent action were associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 381 respondents who rated joint decision speed 1 to 3, 245 (64.30%) had a score of 3 or below. Among the 379 who rated it 4 to 7, 12 (3.17%) had a score of 3 or below, and none of the 26 who rated it 8 to 10 did (chi-square = 335.79, $p < 0.001$, Cramér's V = 0.65).

Decision speed was associated with the score at $r = 0.92$ (95% CI 0.89 to 0.94, $n = 150$ engagements) and with governance in the other four categories at $r = 0.79$. After adjustment for the mean of the other 53 indicators, the association remained at partial $r = 0.23$ ($p = 0.005$).

The association was observed within the canceled engagements at $r = 0.76$ and ranged across the active-stage groups from 0.82 in Build-Out and Transition/Cut-Over to 0.87 in Production. After Holm adjustment, it was significant in 85 of the 86 segment levels tested. Within role it stood at 0.87 among individual contributors and engagement managers and at 0.90 among portfolio directors. Among individual contributors alone, 66.67% (202 of 303) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.69 ($n = 381$), against 5.28 ($n = 405$) among the rest (cluster-robust $p < 0.001$).

Fourteen respondents on 13 engagements wrote about decision speed, and 13 of the answers were negative, 7 from canceled and 6 from active engagements. Three of the 13 described problems raised with no response. The answers set the hour in which an agent acted against the days that two people took to agree on what it had done, and a weekly meeting against a queue the agent worked every hour.

Part 2 (Co-Management, section 5.2) reports the association of decision speed with the four study measures. The pace of governance against the pace of the agents is analyzed as a reason for reported governance weakness under The Fit Between Governance Processes and the Pace of Agent Work.

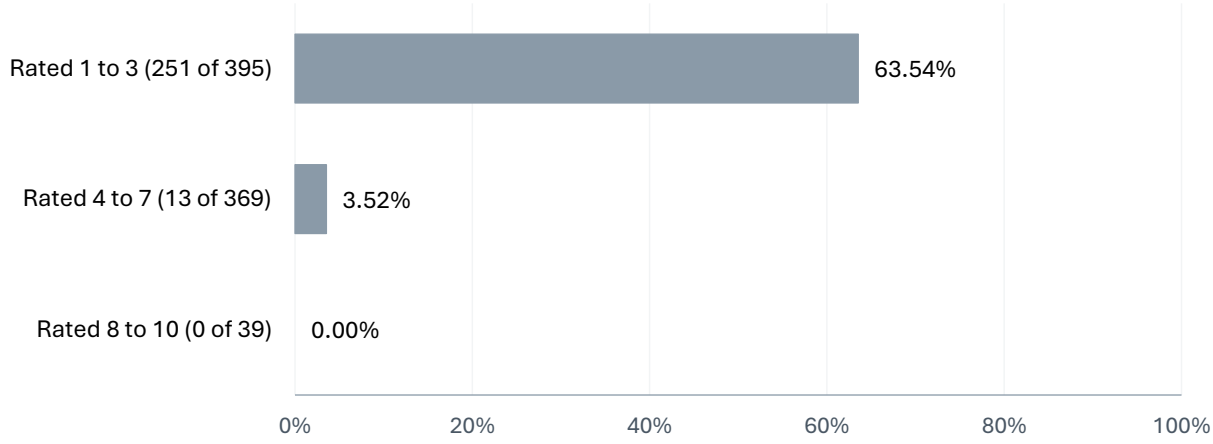
5.3 Fit of the Joint Governance Model to Work Performed by Agents

Indicator C154 · Rate how well today's joint governance model fits work performed by agents rather than work performed by people · Full statistical analysis ›

A joint model poorly fitted to agent-performed work was associated with low scores. The concern is whether authority, review, and issue management are designed for the work now being performed, rather than carried over without adaptation from people-based delivery. The small written base illustrates reported gaps in authority. It does not establish that all arrangements inherited from earlier work are ineffective.

Figure 3.43. Fit of the Joint Governance Model to Work Performed by Agents: share with a score of 3 or below by indicator rating.

A joint model poorly fitted to agent-performed work was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 395 respondents who rated the fit of the joint model to agent work 1 to 3, 251 (63.54%) had a score of 3 or below. Among the 369 who rated it 4 to 7, 13 (3.52%) had a score of 3 or below, and none of the 39 who rated it 8 to 10 did (chi-square = 331.52, $p < 0.001$, Cramér's V = 0.64).

The fit of the model was associated with the score at $r = 0.92$ (95% CI 0.89 to 0.94, n = 150 engagements) and with governance in the other four categories at $r = 0.78$. After adjustment for the mean of the other 53 indicators, the association remained at partial $r = 0.26$ ($p = 0.001$), the fourth-strongest of the 20.

The association was observed within the canceled engagements at $r = 0.77$ and across the active stages at 0.81 to 0.85, and after Holm adjustment it held in 86 of the 87 segment levels tested. Within role it stood at 0.87 among individual contributors and engagement managers

and at 0.80 among portfolio directors. Among individual contributors alone, 65.83% (210 of 319) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.73 (n = 395), against 5.24 (n = 408) among the rest (cluster-robust $p < 0.001$).

Ten respondents on 9 engagements wrote about the fit of the model, and 9 of the answers were negative, 5 from canceled and 4 from active engagements. Three of the 9 described one of the 26 forms of a control bypassed or not working. The answers described leads on both sides who held no authority to decide, and a model that worked issues as one team while keeping no joint record of what the agents had changed.

> *"To be fair to the leads on both sides, none of them were given any power to decide. They sit in the same queue we do... every exception we raise goes up and stops somewhere above them. I don't think they wanted it this way, they simply had nothing to work with."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Corporate and back-office functions

> *"Issues do get worked as one team, that part functions. The rest is built for people doing the work. There is no joint record of what the agents changed between reviews, so when a forecast number moves we cannot say which side moved it. It gets discussed and then it repeats next cycle"*
>
> **ENGAGEMENT MANAGER**
>
> KPO, Corporate and back-office functions

Part 2 (Co-Management, section 5.3) reports the association of the fit of the joint model with the four study measures. The inheritance of governance functions designed for people is analyzed as a reason for reported governance weakness under The Fit Between Governance Processes and the Pace of Agent Work.

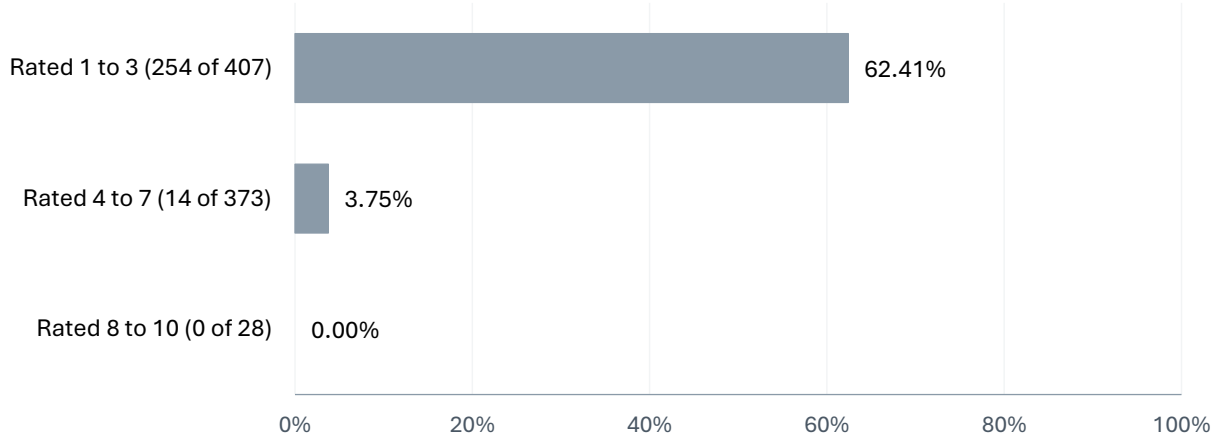
5.4 Readiness of the Co-Management Model for Expected Agent Autonomy

Indicator C155 · Rate how ready the joint management model is for the level of agent autonomy you expect in this engagement within two years · Full statistical analysis ›

Reported readiness for expected agent autonomy was strongly associated with the wider score. This places the anticipated expansion of autonomous work within the governance question: the joint model must be assessed against the authority agents are expected to exercise. The measure concerns respondents' expectations within two years, not observed future performance. Its high correlation is evidence of aligned assessments, not proof that future autonomy will be governed successfully.

Figure 3.44. Readiness of the Co-Management Model for Expected Agent Autonomy: share with a score of 3 or below by indicator rating.

Reported readiness for expected agent autonomy was strongly associated with the wider score.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

The association with the score stood at r = 0.93 (95% CI 0.90 to 0.95, n = 150 engagements). Readiness was associated with governance in the other four categories at r = 0.80, the highest of the 20, and after adjustment for the mean of the other 53 indicators the association remained at partial r = 0.22 (p = 0.008).

Scores were 3 or below for 254 of 407 (62.41%) at indicator ratings 1–3. The corresponding counts were 14 of 373 (3.75%) at indicator ratings 4–7 and 0 of 28 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 316.46, p < 0.001, Cramér's V = 0.63. Respondents who rated it distressed recorded a score of 2.70 (n = 407), against 5.30 (n = 401) among the rest (cluster-robust p < 0.001).

The association was observed within the canceled engagements at r = 0.76 and ranged across the active-stage groups from 0.83 in Build-Out to 0.88 in Production. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.88 among individual contributors, 0.87 among engagement managers, and 0.82 among portfolio directors. Among individual contributors alone, 63.20% (213 of 337) of those who rated it 1 to 3 had a score of 3 or below.

Three respondents on 3 active engagements wrote about readiness for expected autonomy, and two of the answers were negative. Neither negative answer described one of the 26 forms of a control bypassed or not working. One individual contributor judged that the joint model would not hold at the autonomy expected within two years, and a portfolio director on another engagement reported that the conversation on which alert types an agent might close alone had been deferred twice.

Part 2 (Co-Management, section 5.4) reports the association of readiness with the four study measures.

Financial Institution Readiness and the Governance of Agentic Work

Financial institution readiness was closely associated with the reported control environment. Low ratings were common across the institution's data governance, management, access, accountability, and oversight responsibilities. The governance, oversight, and control score links those responsibilities with controls elsewhere in the engagement. Conditional cross-category results were less consistent, with sign changes under alternative comparison means. These qualifications limit distinct-effect claims without removing the institution's substantive responsibilities from the analysis.

Readiness indicators correlated with the score excluding the tested indicator at r = 0.75–0.80, and with governance in the other four categories at r = 0.72–0.76 (n = 150). Five adjusted cross-category estimates were small, negative, and nonsignificant, from −0.15 to −0.11. Response speed was negative and nominally significant in the main model but sensitive to category removal. Data governance had a positive estimate of 0.22. Within canceled engagements, item-to-score correlations were 0.54–0.62. These subgroup comparisons do not establish a change caused by cancellation.

The readiness ratings separated failed from higher composite scores less sharply than the management ratings of the two firms. Among respondents who rated a readiness indicator 1 to 3, 54.26% to 57.31% had a score of 3 or below. Among those who rated one 4 to 7, 7.99% to 10.00% still did, against 2.65% to 5.30% for the co-management indicators. Table 3.9 sets out the seven indicators against these measures.

Table 3.9. The seven financial institution readiness indicators against reported governance weakness across the engagement

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Maturity of the Financial Institution's Data Governance	229 of 401 (57.11%)	0.75 (0.67 to 0.82)	0.74
Effectiveness of the Financial Institution's Management of Its Side	227 of 398 (57.04%)	0.78 (0.71 to 0.84)	0.73
Speed of the Financial Institution's Response to Issues Raised	229 of 404 (56.68%)	0.79 (0.72 to 0.84)	0.75

Table 3.9. The seven financial institution readiness indicators against reported governance weakness across the engagement (continued)

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Assignment of Ownership and Accountability for AI	232 of 417 (55.64%)	0.78 (0.71 to 0.83)	0.73
Capacity to Govern and Oversee Agent Risk	239 of 417 (57.31%)	0.79 (0.72 to 0.84)	0.74
Change Control for the Changes the Engagement Requires	223 of 411 (54.26%)	0.80 (0.74 to 0.85)	0.76
Ability to Provision Access and Identities for Agents	232 of 410 (56.59%)	0.77 (0.69 to 0.83)	0.72

Indicator	Adjusted partial r (p)	Negative written answers
Maturity of the Financial Institution's Data Governance	0.22 (p = 0.006)	152 of 172
Effectiveness of the Financial Institution's Management of Its Side	−0.13 (p = 0.119)	184 of 196
Speed of the Financial Institution's Response to Issues Raised	−0.18 (p = 0.033)	437 of 498
Assignment of Ownership and Accountability for AI	−0.11 (p = 0.201)	41 of 46
Capacity to Govern and Oversee Agent Risk	−0.15 (p = 0.068)	37 of 45
Change Control for the Changes the Engagement Requires	−0.14 (p = 0.088)	69 of 85
Ability to Provision Access and Identities for Agents	−0.12 (p = 0.133)	403 of 437

Note: Rating-band and item-to-score results use the score excluding the tested indicator. Cross-category results exclude its entire category and use 150 engagements. Partial r accounts for the mean of the 53 indicators outside the score. Adjusted-correlation p values are nominal. The ordinary respondent-level band tests do not account for clustering within engagements, including where Monte Carlo calculation handles sparse cells. Their fractions remain descriptive. Zero observations are not proof of no risk.

> *"What the scale does not capture is that the finance team we were given had no mandate. Their data itself was, I would say, in reasonable order; it was the absence of anyone on their side empowered to decide that ended this."*
>
> **PORTFOLIO DIRECTOR**
>
> KPO, Corporate and back-office functions

1 Financial-Institution-Provided Data

Financial-institution-provided data connect governance of inputs with manual correction inside delivery. The rated data-governance indicator and written accounts of absent validation answer related but different questions. One describes perceived capability, while the other identifies reported occasions when input checks were missing before handover. Reading them together clarifies the control boundary without treating the accounts as an independently verified input-error rate.

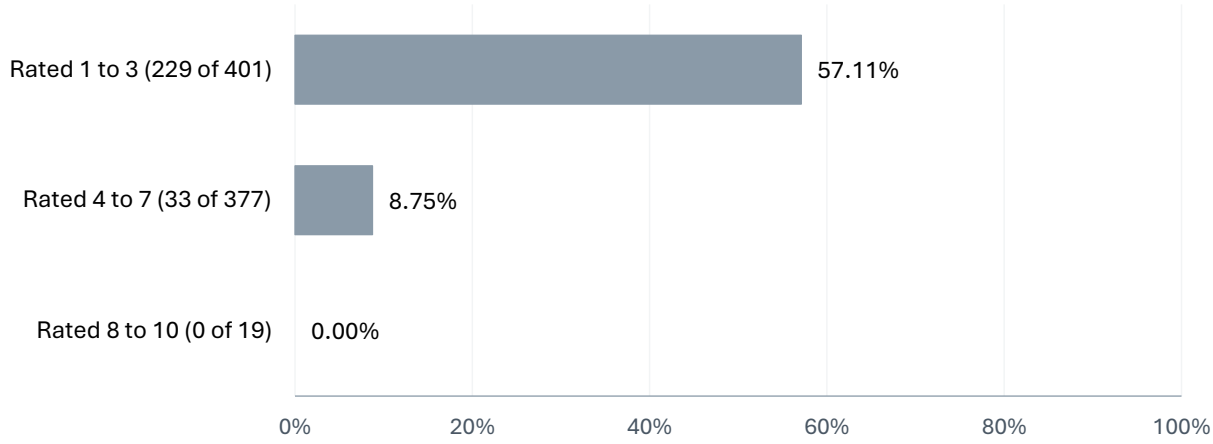
1.1 Maturity of the Financial Institution's Data Governance

Indicator C097 · Rate the maturity of the customer's data governance over the data this engagement uses · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 1.3 · Full statistical analysis ›

Weak data-governance ratings were associated with a weaker wider control environment, with a positive adjusted cross-category estimate in the main model. The financial institution's control over the data used by the engagement provides the substantive connection. Data definitions, accountability, and preparation are part of the conditions under which the shared work operates. The adjusted estimate is model dependent and its significance must be read under the stated test family.

Figure 3.45. Maturity of the Financial Institution's Data Governance: share with a score of 3 or below by indicator rating.

Weak data-governance ratings were associated with a weaker wider control environment, with a positive adjusted cross-category estimate in the main model.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Data governance was associated with the score computed without it at r = 0.75 (95% CI 0.67 to 0.82, n = 150 engagements), the lowest of the 47 governance indicators. It was associated with governance in the other four categories at r = 0.74, and after that adjustment the association remained at partial r = 0.22 (p = 0.006).

Of the 401 respondents who rated data governance 1 to 3, 229 (57.11%) had a score of 3 or below. Among the 377 who rated it 4 to 7, 33 (8.75%) still had a score of 3 or below, and none of the 19 who rated it 8 to 10 did (chi-square = 215.42, p < 0.001, Cramér's V = 0.52). Respondents who rated it distressed recorded a score of 2.89 (n = 401), against 5.11 (n = 396) among the rest (cluster-robust p < 0.001).

The association was the weakest of the 47 within the canceled engagements, at r = 0.54, against 0.64 to 0.71 across the active stages. After Holm adjustment, it was significant in 84 of the 87 segment levels tested. Within role it stood at 0.74 among individual contributors and 0.71 among engagement managers and fell to 0.51 among the 31 portfolio directors, the lowest within-role value of any governance indicator. Among individual contributors alone, 59.38% (193 of 325) of those who rated it 1 to 3 had a score of 3 or below.

Data governance tracked the manual alteration of financial-institution-provided data across its rating bands. Of the respondents who rated data governance 1 to 3, 75.19% (300 of 399) altered the data, against 24.39% (90 of 369) at ratings of 4 to 7 and 1 of 19 at ratings of 8 to 10 (chi-square p < 0.001). Part 5 (Leading Dynamics of Manual Alteration, section 1.3) reports that comparison in full.

Data governance drew 172 written answers from 155 respondents on 99 engagements, and 152 were negative, 77 from canceled and 75 from active engagements. Of the 152, 57 (37.50%) described one of the 26 forms of a control bypassed or not working, and 50 of the 57 described data handed over with no validation by the financial institution. Respondents described fields with no owner of their meaning, classifications that changed from month to month with no approver, and governance rules written after the work had gone live.

> *"vendor master made me angry from first week... same supplier written three different ways in three systems, legal name in one, short name in other. nobody owns which is correct"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Corporate and back-office functions

Part 2 (Financial Institution Readiness, section 1.4) reports the association of data governance with the four study measures.

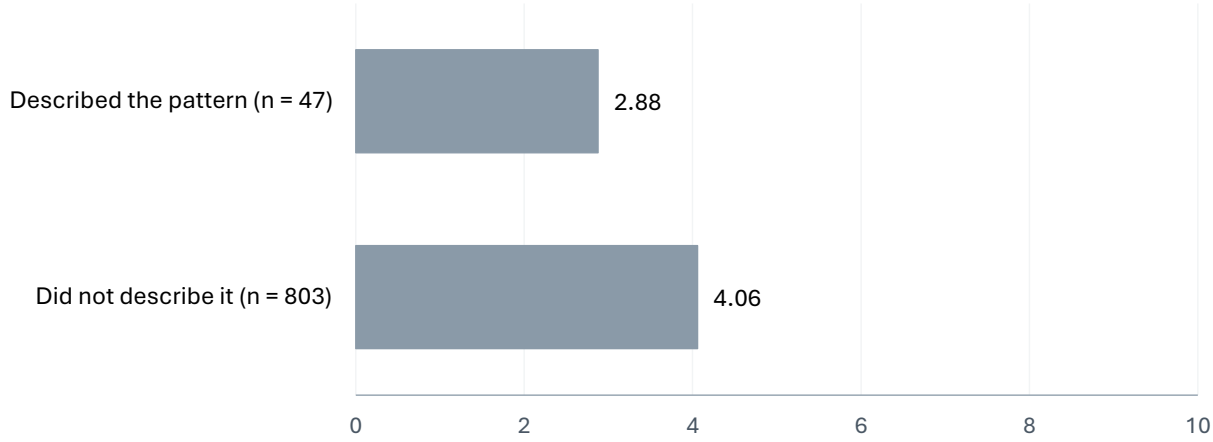
1.2 Absence of Validation by the Financial Institution Before Handover

Written pattern · All written answers to the open questions · 50 answers from 47 respondents on 37 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 1.4

Accounts of data handed over without validation came from respondents with lower mean scores. This connects a specific input-control gap with the broader assessment of governance. The operational concern is that checking and correction may occur after handover through the delivery workforce, with responsibilities less visible than a validated input process would make them. The written pattern identifies reported experience, not the incidence of all unvalidated data deliveries.

Figure 3.46. Absence of Validation by the Financial Institution Before Handover: descriptive mean score comparison.

Accounts of data handed over without validation came from respondents with lower mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 47 respondents who described it recorded a score of 2.88, against 4.06 among the other 803 (cluster-robust $p < 0.001$). Among individual contributors alone, the 47 who described it recorded 2.88 against 3.94 among the other 620 (cluster-robust $p < 0.001$).

The engagement share describing unvalidated data correlated with the score at $r = -0.24$ ($p = 0.004$, $n = 150$). After adjustment for the comparison mean, partial r was 0.01 (nominal $p = 0.922$). The adjusted analysis did not detect a separate association. It does not establish that absent validation has no operational relationship with governance.

The pattern differed across engagement status and gathered where the composite score was low. It was described by 2.86% of respondents in Build-Out, 4.00% in Transition/Cut-Over, 6.80% in Production, and 7.76% on canceled engagements, and by 9.09% in the lowest indicator quintile against 1.88% in the highest. By the score, it was described by 9.19% (26 of 283) of respondents at 3 or below, 3.98% (21 of 528) above 3 to 7, and none of the 39 above 7. All 47 who described it were individual contributors.

The share describing the pattern differed after Holm adjustment by management's understanding of the governing obligations, by the financial institution's discussion of them, by the manual alteration methods used, by disclosure to the financial institution, and by intent to leave. It differed by none of the structural attributes of the engagement.

The 50 answers came from all written answers to the open questions, and 26 of them were written on canceled engagements. Part 5 reports the same 50 answers from the same 47 respondents on 37 engagements, 49 of them answers on the manual alteration of data, so the two counts describe one body of testimony. Respondents described correcting the data on every file before the agent could run, tickets to the financial institution that went unanswered, and a delivery team that had become the only validation the data received.

> *"the problem is nothing is validated on their side before it reaches us, so i fix data myself every time, further raising ticket does nothing at all, nobody replies to it"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Financial crime and client onboarding

Part 5 (Leading Dynamics of Manual Alteration, section 1.4) analyzes the pattern as a factor in manual alteration.

2 Management of the Financial Institution's Side of the Engagement

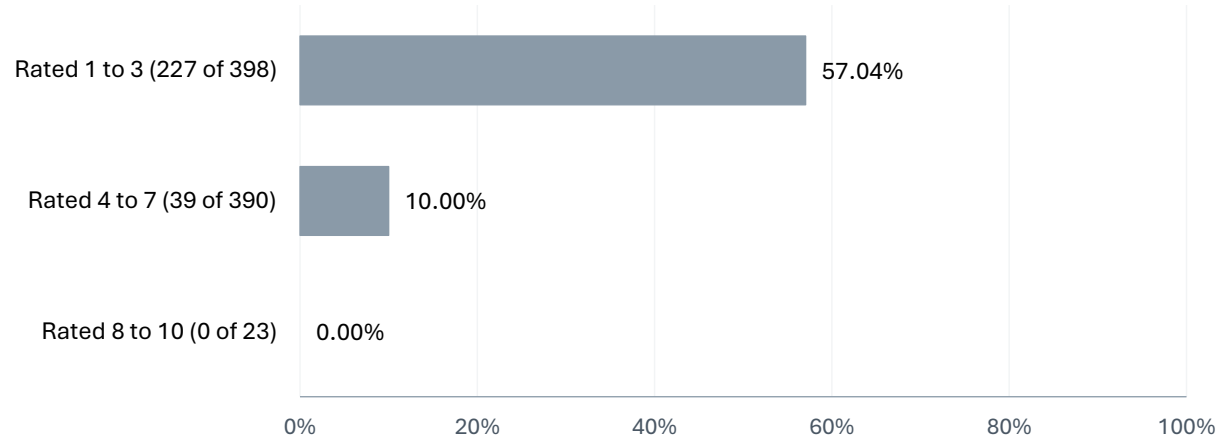
Management of the financial institution's own work and its response to issues were associated with the wider control environment. These responsibilities concern actions that the vendor cannot complete solely through its own delivery process. The adjustment results show model dependence, particularly for response speed. The operational interpretation rests on the reported responsibilities and evidence, rather than a claim that the coefficients isolate their causal effects.

2.1 Effectiveness of the Financial Institution's Management of Its Side

Indicator C105 · Rate the effectiveness of the customer's management of its side of this engagement · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 3.2 · Full statistical analysis ›

Weak management of the financial institution's side was associated with low composite scores, while the adjusted cross-category analysis did not detect a clear association. This indicator concerns the institution's ability to perform its own engagement responsibilities. It extends the governance question beyond vendor delivery, because shared work also depends on decisions and actions inside the institution. The conditional result does not establish that those responsibilities have no distinct operational role.

Figure 3.47. Effectiveness of the Financial Institution's Management of Its Side: share with a score of 3 or below by indicator rating.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 398 respondents who rated it 1 to 3, 227 had a score of 3 or below. Among the 390 who rated it 4 to 7, 39 (10.00%) still did, the highest share at those ratings of the seven readiness indicators, and none of the 23 who rated it 8 to 10 did (chi-square = 209.26, $p < 0.001$, Cramér's V = 0.51).

The indicator was associated with the score computed without it at $r = 0.78$ (95% CI 0.71 to 0.84, n = 150 engagements) and with governance in the other four categories at $r = 0.73$. After adjustment for the mean of the other 53 indicators, the partial r was −0.13 and was not significant ($p = 0.119$). Respondents who rated it distressed recorded a score of 2.87 (n = 398), against 5.09 (n = 413) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at $r = 0.58$ and across the active stages at 0.68 to 0.79, and after Holm adjustment it held in 85 of the 87 segment levels tested. Within role it stood at 0.77 among individual contributors, 0.73 among engagement managers, and 0.75 among portfolio directors. Among individual contributors alone, 59.44% (192 of 323) of those who rated it 1 to 3 had a score of 3 or below.

The indicator drew 196 written answers from 173 respondents on 104 engagements, and 184 of them were negative, the most uniformly negative testimony among the governance indicators with 100 or more answers. Of the 184 negative answers, 105 came from canceled and 79 from active engagements, and 48 (26.09%) described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (21), undefined handoffs between the financial institution's teams (14), and raising problems seen as futile (9). Respondents described a financial institution with no single owner of the agent work, questions that passed among three functions, and staff who held no authority to approve anything.

> *"My honest frustration — and it is the only one worth writing down — is that there is no single owner on their side for the agent work. Procurement operations, their vendor risk function and the technology group each hold a piece, so a question about whether an agent may close a supplier onboarding case without a human goes round all three and comes back a fortnight later, usually unanswered. Everything else about them is genuinely solid."*
>
> **ENGAGEMENT MANAGER**
>
> BPO, Corporate and back-office functions

> *"The worst of it is nobody over there is actually running their end. I can't get into the Custodian portal half the time and break files land in three different layouts. But their people know the tech fine. They just aren't set up to run this."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Trade, settlement and reconciliation

Part 2 (Financial Institution Readiness, section 5.1) reports the association of the indicator with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 3.2) its association with manual alteration.

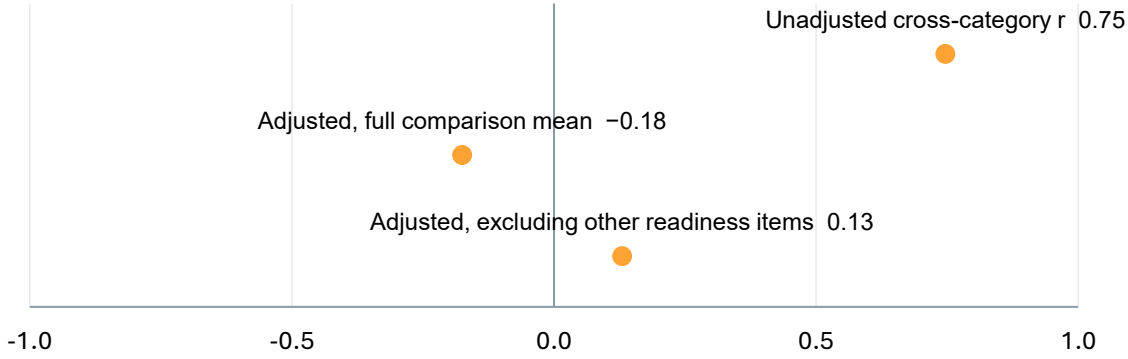
2.2 Speed of the Financial Institution's Response to Issues Raised

Indicator C109 · Rate how quickly the customer responds when this engagement raises issues that need customer action · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 3.1 · Full statistical analysis ›

The financial institution's response to issues was closely associated with the wider control environment and drew extensive written discussion. Issues requiring financial institution action need to be resolved while work continues. Its adjusted coefficient changed sign when the institution's other readiness indicators were removed from the comparison mean. That sensitivity limits a distinct-effect interpretation and gives the operating accounts more explanatory value than a ranking of comment volume.

Figure 3.48. Speed of the Financial Institution's Response to Issues Raised: observed association and adjustment sensitivity.

The financial institution's response to issues was closely associated with the wider control environment and drew extensive written discussion.



Comparison	Estimate and analysis base
Unadjusted cross-category r	r = 0.75 150 engagements
Adjusted, full comparison mean	r = −0.18 150 engagements
Adjusted, excluding other readiness items	r = 0.13 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

Correlations use 150 engagements. The alternative mean removes the stated category items. Estimates are descriptive model comparisons, with nominal inference in the text.

Accounts of response speed included unresolved decisions while delivery continued, linking the indicator to the availability of timely financial institution action. It was addressed in 498 written answers from 382 respondents on 142 engagements, more than any other governance indicator.

The indicator was associated with the score computed without it at r = 0.79 (95% CI 0.72 to 0.84, n = 150 engagements) and with governance in the other four categories at r = 0.75. After adjustment for the mean of the other 53 indicators, the partial r was −0.18 (p = 0.033), the one nominally significant inverse association among the readiness indicators. The 53 include the financial institution's other readiness indicators. Leaving them out of the control moved the partial r to 0.13 (p = 0.111), while leaving out the vendor capability block left it at −0.35 (p < 0.001), the AI technology block at −0.04 (p = 0.594), and the third-party block at −0.15 (p = 0.065). These alternative specifications show sensitivity to the comparison measure. They do not establish that response speed has no relationship of its own.

Scores were 3 or below for 229 of 404 (56.68%) at indicator ratings 1–3. The corresponding counts were 30 of 367 (8.17%) at indicator ratings 4–7 and 0 of 18 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 214.22, p < 0.001, Cramér's V = 0.52. Respondents who rated it distressed recorded a score of 2.91 (n = 404), against 5.18 (n = 385) among the rest (cluster-robust p < 0.001).

The association was observed within the canceled engagements at r = 0.61, the second-strongest of the seven readiness indicators, and across the active stages at 0.65 to 0.76. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.77 among individual contributors, 0.73 among engagement managers, and 0.63 among portfolio directors. Among individual contributors alone, 57.82% (196 of 339) of those who rated it 1 to 3 had a score of 3 or below.

Of the 498 written answers, 437 were negative, 217 from canceled and 220 from active engagements. Of the 437, 181 (41.42%) described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (120), raising problems seen as futile (35), and undefined handoffs between the financial institution's teams (21). Respondents described requests for a supplier list that went unanswered for weeks, and exceptions held open late into the evening while the financial institution did not say which record was correct.

> *“i felt like i was working for nobody. we asked for the vendor master and the supplier list for weeks, no reply. two months on that desk and their side never turned up once”*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Corporate and back-office functions

> *This is not my job but I am the one doing it. Their side does not answer… I sit with PO exceptions until late and nobody says which record is correct. My team is tired, two already left. Then they ask why the queue is old"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Corporate and back-office functions

Part 2 (Financial Institution Readiness, section 5.2) reports the association of response speed with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 3.1) its association with the manual alteration of processes.

“

Question about change control assumes there is change control to rate. For payroll there is none we can point to. We raised a form in March for adding one benefit code and it is still open, and meanwhile the Cycle runs anyway. Same for data question, score of 1 does not separate file that is late from file that is wrong, and we get both, from hris and from the tax side, with definitions that do not agree between them. Their people who own AI are at least named, that part is clear enough. For the rest the scale is too narrow.”

ENGAGEMENT MANAGER

BPO, Corporate and back-office functions

3 Governance of Agentic AI Within the Financial Institution

Internal ownership, oversight capacity, and change control each connect the financial institution's organization with governance of agent work. The indicators distinguish naming responsibility, having the capacity to exercise it, and handling required changes. Their similar score relationships should not obscure those separate functions. The adjusted results qualify a claim of a distinct statistical association for each control.

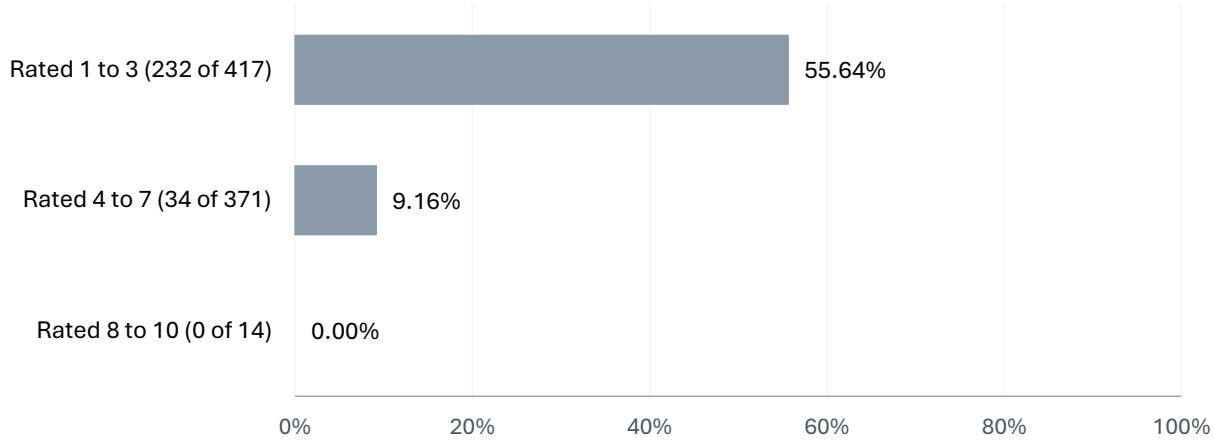
3.1 Assignment of Ownership and Accountability for AI

Indicator C112 · Rate how clearly ownership and accountability for AI are assigned inside the customer organization · Full statistical analysis ›

Unclear assignment of AI ownership and accountability inside the financial institution was associated with low scores. The indicator addresses an internal responsibility boundary that joint vendor-institution procedures cannot settle by themselves. It asks whether the institution has identified who owns AI-related decisions and consequences within its organization. The findings concern reported clarity, with no independent organizational audit of those assignments.

Figure 3.49. Assignment of Ownership and Accountability for AI: share with a score of 3 or below by indicator rating.

Unclear assignment of AI ownership and accountability inside the financial institution was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 417 respondents who rated ownership 1 to 3, 232 had a score of 3 or below. Among the 371 who rated it 4 to 7, 34 (9.16%) still did, and none of the 14 who rated it 8 to 10 did (Monte Carlo chi-square = 198.34, $p < 0.001$, Cramér's V = 0.50).

Ownership was associated with the score computed without it at $r = 0.78$ (95% CI 0.71 to 0.83, n = 150 engagements) and with governance in the other four categories at $r = 0.73$. After adjustment for the mean of the other 53 indicators, the partial r was −0.11 and was not significant ($p = 0.201$). Respondents who rated it distressed recorded a score of 2.95 (n = 417), against 5.12 (n = 385) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at $r = 0.56$ and across the active stages at 0.67 to 0.74, and after Holm adjustment it held in 85 of the 87 segment levels tested. Within role it stood at 0.76 among individual contributors, 0.74 among engagement managers, and 0.65 among portfolio directors. Among individual contributors alone, 57.60% (197 of 342) of those who rated it 1 to 3 had a score of 3 or below.

Forty-six respondents on 40 engagements wrote about ownership of AI, one answer each, and 41 of the answers were negative, 18 from canceled and 23 from active engagements. Of the 41, 10 described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (7) and undefined handoffs between the financial institution's teams (5). Those five answers described ownership divided among functions rather than absent, with risk, technology, and operations teams each assigning the agent to another.

> *"There was no named owner on the bank side for the screening model. Compliance, IT and the AML operations team each said it belonged to the other."*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Financial crime and client onboarding

> *"We spent most of the POC working around a surveillance stack that predates all of us, and their engineers are as frustrated with it as we are. The harder part is that no one on their side owns the AI decisions, so approvals wander"*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Regulatory, audit, tax and legal

Part 2 (Financial Institution Readiness, section 6.1) reports the association of ownership with the four study measures.

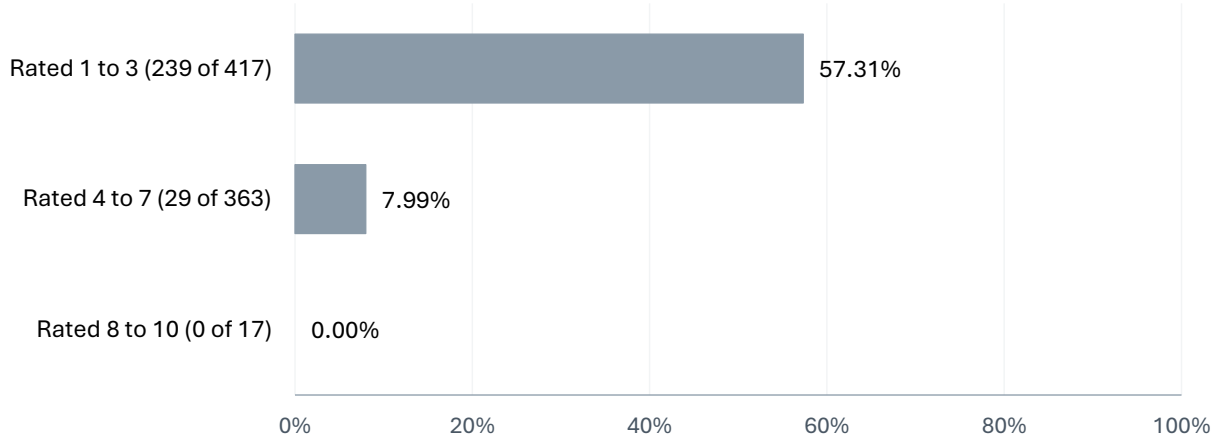
3.2 Capacity to Govern and Oversee Agent Risk

Indicator C111 · Rate the customer's capacity to govern and oversee the risks of AI agents doing this work · Full statistical analysis ›

Limited capacity to oversee agent risk was associated with low scores. This concerns whether the financial institution can examine and govern the risks introduced by agent work, in addition to meeting its ordinary management responsibilities. The main adjusted model did not detect a clear cross-category association. That result qualifies the statistical claim while preserving the distinction between nominal responsibility and reported capacity to exercise oversight.

Figure 3.50. Capacity to Govern and Oversee Agent Risk: share with a score of 3 or below by indicator rating.

Limited capacity to oversee agent risk was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 417 respondents who rated oversight capacity 1 to 3, 239 had a score of 3 or below. Among the 363 who rated it 4 to 7, 29 (7.99%) still did, the lowest such share of the seven, and none of the 17 who rated it 8 to 10 did (chi-square = 220.35, $p < 0.001$, Cramér's V = 0.53).

Oversight capacity was associated with the score computed without it at $r = 0.79$ (95% CI 0.72 to 0.84, n = 150 engagements) and with governance in the other four categories at $r = 0.74$. After adjustment for the mean of the other 53 indicators, the partial r was −0.15 and was not significant ($p = 0.068$). Respondents who rated it distressed recorded a score of 2.95 (n = 417), against 5.16 (n = 380) among the rest (cluster-robust $p < 0.001$).

The association was weakest in Build-Out, at $r = 0.60$, against 0.79 in Transition/Cut-Over and 0.73 in Production, and held within the canceled engagements at 0.62, the strongest of the seven. After Holm adjustment, it was significant in 84 of the 87 segment levels tested. Within

role it stood at 0.78 among individual contributors, 0.74 among engagement managers, and 0.71 among portfolio directors. Among individual contributors alone, 60.00% (201 of 335) of those who rated it 1 to 3 had a score of 3 or below.

Forty-five respondents on 38 engagements wrote about oversight capacity, and 37 of the answers were negative, 19 from canceled and 18 from active engagements. Only 3 of the 37 described one of the 26 forms of a control bypassed or not working. Respondents described audit frameworks that did not exist for the agents and model-risk governance built after launch. A portfolio director on a canceled payments engagement reported that the financial institution's AI oversight forum had been in place before the vendor arrived, while ownership of the data it oversaw never settled.

> *"The client was not an unserious counterparty. On governance they were ahead of most acquirers we have dealt with - their AI oversight forum was standing before we arrived, which was not true of two other payments accounts I have been near - and their process documentation was usable, which is rarer still. What did not move in three months was ownership of the transaction data itself. Nobody on their side could tell us definitively who signs off a change to a merchant reference field, and so the change went to a forum that met when it met. Data governance is the item I marked hardest and I would mark it the same today. The rest of it, access and formats, is ordinary and fixable inside a quarter. That one is not."*
>
> **PORTFOLIO DIRECTOR**
>
> ITO, Technology engineering and operations

Part 2 (Financial Institution Readiness, section 6.2) reports the association of oversight capacity with the four study measures.

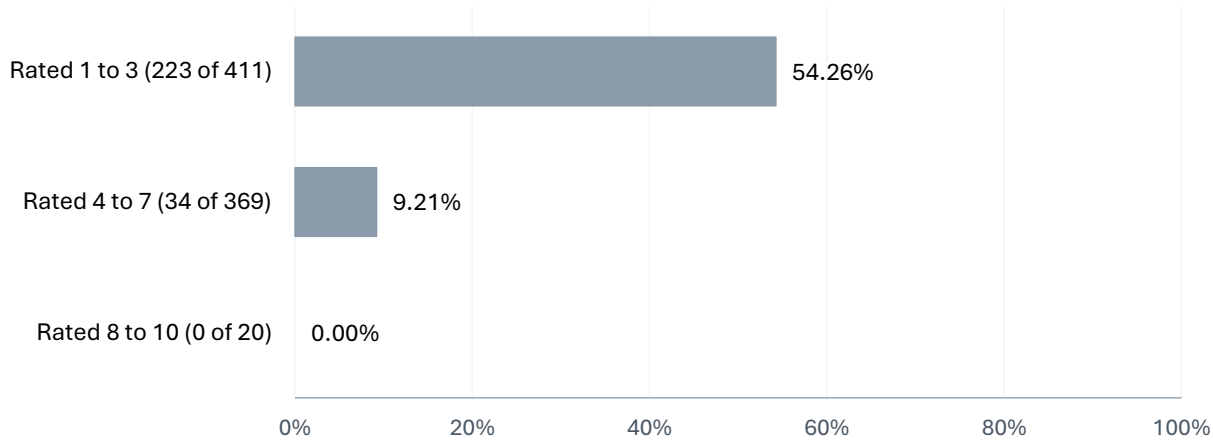
3.3 Change Control for the Changes the Engagement Requires

Indicator C110 · Rate how well the customer's change control process handles the changes this engagement requires · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 3.4 · Full statistical analysis ›

Weak financial institution change-control ratings were associated with a weaker reported control environment. This concerns changes the institution must handle on its own side, distinct from the joint process used between the firms. The main adjusted cross-category estimate was negative and not clearly distinguishable from zero. The substantive interpretation remains a reported readiness constraint, without treating the coefficient as evidence that better change control worsens governance.

Figure 3.51. Change Control for the Changes the Engagement Requires: share with a score of 3 or below by indicator rating.

Weak financial institution change-control ratings were associated with a weaker reported control environment.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Change control was associated with the score computed without it at r = 0.80 (95% CI 0.74 to 0.85, n = 150 engagements), the highest of the seven. After adjustment for the mean of the other 53 indicators, the partial r was −0.14 and was not significant (p = 0.088).

Of the 411 respondents who rated change control 1 to 3, 223 had a score of 3 or below. Among the 369 who rated it 4 to 7, 34 (9.21%) still did, and none of the 20 who rated it 8 to 10 did (chi-square = 190.63, p < 0.001, Cramér's V = 0.49, the smallest conditional share of the 47). Respondents who rated it distressed recorded a score of 2.93 (n = 411), against 5.16 (n = 389) among the rest (cluster-robust p < 0.001).

The association fell further on cancellation than for any other readiness indicator, from 0.63 to 0.77 across the active stages to 0.55 within the canceled engagements. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.77 among individual contributors, 0.76 among engagement managers, and 0.72 among portfolio directors. Among individual contributors alone, 56.63% (188 of 332) of those who rated it 1 to 3 had a score of 3 or below.

Change control drew 85 written answers from 80 respondents on 64 engagements, and 69 were negative, 33 from canceled and 36 from active engagements. Of the 69, 15 (21.74%) described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (7) and habit in place of a control (3). Respondents described a change board that sat on fixed days or once a month against agents that changed continuously, and a form to add one benefit code that was still open while the payroll cycle ran on.

> *“Their change board sits once a month. Nothing gets approved between… And they want same desk process as before, only with agent doing it.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

Part 2 (Financial Institution Readiness, section 6.3) reports the association of change control with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 3.4) its association with manual alteration.

4 Provisioning of System Access and Agent Identities

Provisioning system access and agent identities was associated with the wider control environment and was repeatedly described in accounts of manual work. This responsibility connects authorized access with the ability to operate the intended workflow. The finding concerns reported provisioning capability and does not independently verify permissions, identity security, or the reason for each manual intervention.

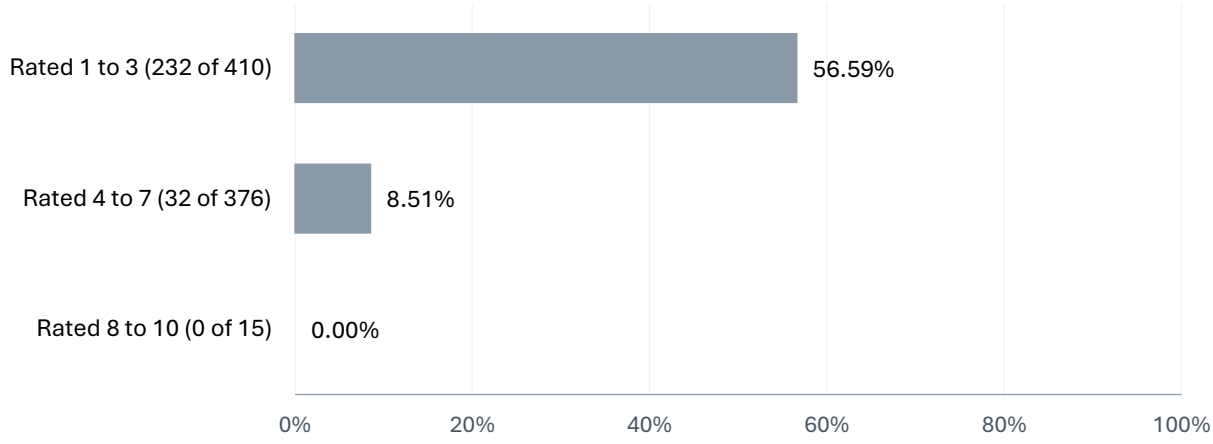
4.1 Ability to Provision Access and Identities for Agents

Indicator C108 · Rate the customer's ability to provision the system access and identities that AI agents require · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 2.1 · Full statistical analysis ›

Difficulty provisioning agent access and identities was associated with low scores. Access is a governance issue as well as a delivery dependency, because the institution must supply the identities and permissions through which agent work is performed. The written accounts connect this responsibility with manual handling of work. The findings describe reported ability and experience, without verifying the security or appropriateness of each identity and permission.

Figure 3.52. Ability to Provision Access and Identities for Agents: share with a score of 3 or below by indicator rating.

Difficulty provisioning agent access and identities was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 410 respondents who rated provisioning 1 to 3, 232 had a score of 3 or below. Among the 376 who rated it 4 to 7, 32 (8.51%) still did, and none of the 15 who rated it 8 to 10 did (Monte Carlo chi-square = 212.66, $p < 0.001$, Cramér's V = 0.52).

Provisioning was associated with the score computed without it at $r = 0.77$ (95% CI 0.69 to 0.83, n = 150 engagements) and with governance in the other four categories at $r = 0.72$, the lowest of the seven. After adjustment for the mean of the other 53 indicators, the partial r was −0.12 and was not significant ($p = 0.133$). Respondents who rated it distressed recorded a score of 2.90 (n = 410), against 5.14 (n = 391) among the rest (cluster-robust $p < 0.001$).

The association was weakest in Build-Out, at $r = 0.58$, against 0.77 in Transition/Cut-Over and 0.74 in Production, and it held within the canceled engagements at 0.57. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.76 among individual contributors, 0.74 among engagement managers, and 0.62 among portfolio directors. Among individual contributors alone, 59.88% (197 of 329) of those who rated it 1 to 3 had a score of 3 or below.

Provisioning drew 437 written answers from 348 respondents on 135 engagements, the second-largest written base of the 47 governance indicators, and 403 were negative, 181 from canceled and 222 from active engagements. Of the 403, 67 (16.63%) described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (29), raising problems seen as futile (13), and unclear liability or regulatory accountability (9). The answers recorded the control used outside its design. Agents ran under shared service accounts or under the credentials of named people, credentials lapsed with no person assigned to renew them, and access management treated an agent identity as a new joiner with a line manager and a leaver date.

> *“The difficulty, and it is the only one I would put in front of their leadership, is identity… Every agent we place on the trust administration work needs its own credential on the client's core, and their access management still treats a credential as a person, with a joiner form, a line manager and a leaver date. We have waited eleven weeks for one service account. Everything else on this account I would call reasonable, the process documentation in particular is better than I usually see… but access is the item that quietly sets our timeline.”*
>
> **PORTFOLIO DIRECTOR**
>
> BPO, Corporate and back-office functions

> *Their merchant settlement files arrive in three formats from three platforms, and two of them still come as an SFTP drop with no schema attached. Agent identities were requested in March and half were still pending in July, so the agents run under a shared service account, which their own risk people have flagged. Their managers are reasonable to deal with. The estate underneath them is not."*
>
> **ENGAGEMENT MANAGER**
>
> BPO, Trade, settlement and reconciliation

> *Problem is only access. Agent still dont have own login on core servicing platform so we run half the account amendments under my team id, then audit come back and ask who did it. Data itself is fine mostly"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Payments, cards and disputes

Part 2 (Financial Institution Readiness, section 7.1) reports the association of provisioning with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 2.1) its association with manual alteration.

“

The people on their side are not the problem and I want that on record. Their category managers answer the same day, and when we ask for a change to how the exceptions are routed it goes through their board in a week, which in my experience is fast for a card issuer. What did not move at all in four months is the supplier master. The same vendor sits in there three times with three tax numbers, spend categories are free text in one region and coded in another, and nobody owns that record. We are asking agents to reason over a file that no human has agreed on.”

ENGAGEMENT MANAGER

BPO, Corporate and back-office functions

Vendor Preparation and Accountability in the Control Environment

The vendor's playbook and accountability ratings were associated with governance across the engagement. Both concern the vendor's ability to make agent work governable, through a route into controlled production and responsibility for costly errors. The governance, oversight, and control score summarizes the wider reported environment. Negative adjusted coefficients changed markedly when other vendor capability indicators were removed from the comparison mean, making model sensitivity central to interpretation.

Each indicator was associated with the score computed without it at r = 0.78 and 0.79 and with governance in the other four failure categories at r = 0.78 (n = 150 engagements). After adjustment for the mean of the other 53 indicators, the two ties turned inverse and significant, at partial r = −0.38 and −0.46, the second-strongest and strongest inverse results of the 47 governance indicators. The 53 include the vendor's other 18 capability indicators. With that block omitted from the comparison mean, the two partial r moved to 0.04 and −0.03, while leaving out any other category's block left them between −0.41 and −0.29. The change in estimates shows sensitivity to the vendor capability ratings in the comparison mean. It limits a distinct-effect interpretation of the negative coefficients. Table 3.10 sets out both indicators against these measures.

Table 3.10. The two vendor governance indicators against reported governance weakness across the engagement

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Maturity of the Playbook From Concept to Governed Production	232 of 392 (59.18%)	0.78 (0.71 to 0.84)	0.78
Readiness to Take Accountability for a Costly Agent Error	240 of 373 (64.34%)	0.79 (0.72 to 0.84)	0.78

Indicator	Adjusted partial r (p)	Negative written answers
Maturity of the Playbook From Concept to Governed Production	−0.38 ($p < 0.001$)	186 of 200
Readiness to Take Accountability for a Costly Agent Error	−0.46 ($p < 0.001$)	29 of 38

Note: Rating-band and item-to-score results use the score excluding the tested indicator. Cross-category results exclude its entire category and use 150 engagements. Partial r accounts for the mean of the 53 indicators outside the score. Adjusted-correlation p values are nominal. The ordinary respondent-level band tests do not account for clustering within engagements, including where Monte Carlo calculation handles sparse cells. Their fractions remain descriptive. Zero observations are not proof of no risk.

1 Delivery Playbook, Quality Assurance, and Platforms

The playbook rating connects preparation for governed production with the wider score. Its observed relationship was positive, while the main adjusted result was negative and became near zero under the category-removal comparison. The substantive question remains whether the vendor has a usable process for establishing controls before and during live operation. The sensitivity precludes interpreting the negative coefficient as a benefit of weaker preparation.

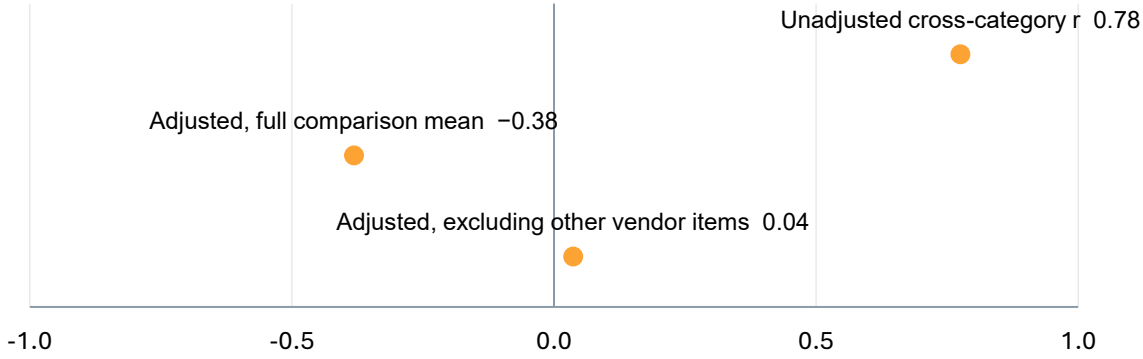
1.1 Maturity of the Playbook From Concept to Governed Production

Indicator C126 · Rate the maturity of your employer's playbook for taking agentic use cases from concept to governed production · Full statistical analysis ›

A less mature vendor playbook was associated with low scores, but its adjusted relationship depended strongly on the comparison measure. The substantive question is how the vendor takes agentic work from concept into governed production. A defined playbook can connect preparation, testing, responsibility, and ongoing operation. The negative conditional coefficient cannot be read as evidence that an immature playbook improves control, given the sensitivity to other vendor capability items.

Figure 3.53. Maturity of the Playbook From Concept to Governed Production: observed association and adjustment sensitivity.

A less mature vendor playbook was associated with low scores, but its adjusted relationship depended strongly on the comparison measure.



Comparison	Estimate and analysis base
Unadjusted cross-category r	r = 0.78 150 engagements
Adjusted, full comparison mean	r = −0.38 150 engagements
Adjusted, excluding other vendor items	r = 0.04 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

Correlations use 150 engagements. The alternative mean removes the stated category items. Estimates are descriptive model comparisons, with nominal inference in the text.

Of the 392 respondents who rated the playbook 1 to 3, 232 had a score of 3 or below. Among the 396 who rated it 4 to 7, 34 (8.59%) still did, and none of the 19 who rated it 8 to 10 did (chi-square = 237.80, $p < 0.001$, Cramér's V = 0.54).

The playbook was associated with the score computed without it at r = 0.78 (95% CI 0.71 to 0.84, n = 150 engagements) and with governance in the other four categories at r = 0.78. After adjustment for the mean of the other 53 indicators, the partial r was −0.38 ($p < 0.001$). The result held when the financial institution readiness block was omitted from the comparison mean (−0.30, $p < 0.001$), when the AI technology block left it (−0.34, $p < 0.001$), and when the third-party block left it (−0.29, $p < 0.001$). It disappeared only when the vendor capability block was omitted from the comparison mean, at 0.04 ($p = 0.660$). Respondents who rated the playbook distressed recorded a score of 2.87 (n = 392), against 5.09 (n = 415) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at r = 0.66 and across the active stages at 0.70 in Build-Out, 0.65 in Transition/Cut-Over, and 0.81 in Production. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.78 among individual contributors, 0.75 among engagement managers, and 0.68 among portfolio directors. Among individual contributors alone, 61.25% (196 of 320) of those who rated it 1 to 3 had a score of 3 or below.

The playbook drew 200 written answers from 184 respondents on 110 engagements, and 186 were negative, 75 from canceled and 111 from active engagements. Of the 186, only 14 (7.53%) described one of the 26 forms of a control bypassed or not working, most often unclear liability or regulatory accountability (7). The answers described a framework that was absent rather than bypassed. Teams were placed on live work with a deck and two days of training, wrote runbooks at night while backlogs grew, and governed the work through habits learned by watching the agents.

> *“There was no playbook. I got two days of training and a deck. And when the fails queue went wrong it came back on me, not on anyone who built it.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Trade, settlement and reconciliation

> *What we have settled into, rather than waiting for the tooling to improve, is a set of habits. Long submissions, and a loss run of forty pages is not unusual, are split into sections before anything is read, because material facts sitting in the middle of a document were being passed over often enough that we stopped trusting a single pass. Each account is worked in its own workspace, opened fresh and closed at the end of the day, so that nothing from one carrier's file can find its way into another's. And where an agent has to touch a system that writes rather than reads, one of my seniors approves the action by hand. None of this is in the playbook. It is what four months of watching the thing has taught us, and hence my middling scores rather than poor ones."*
>
> **ENGAGEMENT MANAGER**
>
> KPO, Credit, underwriting and suitability

Part 2 (Vendor Capability, section 6.1) reports the association of the playbook with the four study measures.

> *The vendor's incident log classified access errors as customer-caused whenever the underlying policy data came from the customer's systems, even when the agent's own logic misapplied that policy."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

2 Accountability for Agent Error

Vendor accountability for agent error was associated with the wider score, but its conditional estimate depended on the adjustment measure. This control concerns responsibility for the consequences of agent action, beyond routine task completion. The evidence supports examining that responsibility while resisting a causal interpretation of the negative adjusted coefficient.

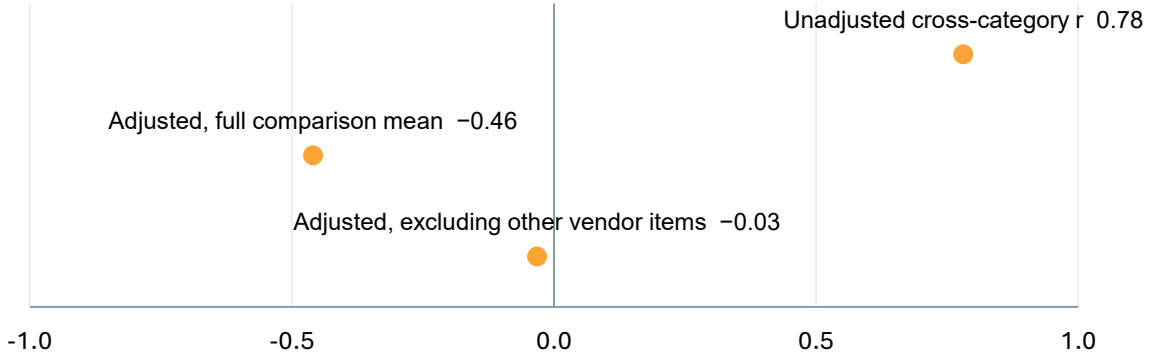
2.1 Readiness to Take Accountability for a Costly Agent Error

Indicator C130 · Rate your employer's readiness to take accountability when an agent it operates makes a costly error · Full statistical analysis ›

Limited vendor readiness to accept accountability for a costly agent error was associated with low scores. This concerns responsibility when the vendor-operated agent causes a material problem, beyond the existence of a delivery process. The observed positive association and the negative adjusted coefficient answer different questions. The latter largely disappeared when the other vendor capability items were removed from the comparison mean, limiting a separate-effect interpretation.

Figure 3.54. Readiness to Take Accountability for a Costly Agent Error: observed association and adjustment sensitivity.

Limited vendor readiness to accept accountability for a costly agent error was associated with low scores.



Comparison	Estimate and analysis base
Unadjusted cross-category r	r = 0.78 150 engagements
Adjusted, full comparison mean	r = −0.46 150 engagements
Adjusted, excluding other vendor items	r = −0.03 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

Correlations use 150 engagements. The alternative mean removes the stated category items. Estimates are descriptive model comparisons, with nominal inference in the text.

Of the 373 respondents who rated the vendor's readiness 1 to 3, 240 had a score of 3 or below, a share within the co-management range. Among the 396 who rated it 4 to 7, 32 (8.08%) still did, and none of the 25 who rated it 8 to 10 did (chi-square = 283.42, $p < 0.001$, Cramér's V = 0.60).

The indicator was associated with the score computed without it at r = 0.79 (95% CI 0.72 to 0.84, n = 150 engagements) and with governance in the other four categories at r = 0.78. After adjustment for the mean of the other 53 indicators, the partial r was −0.46 (p < 0.001). It stayed inverse with the financial institution readiness block omitted from the comparison mean (−0.41), with the AI technology block left out (−0.38), and with the third-party block left out (−0.32), each at p < 0.001. With the vendor capability block left out it stood at −0.03 (p = 0.696). Respondents who rated it distressed recorded a score of 2.75 (n = 373), against 5.05 (n = 421) among the rest (cluster-robust p < 0.001).

The association was observed within the canceled engagements at r = 0.69 and across the active stages at 0.71 in Build-Out, 0.66 in Transition/Cut-Over, and 0.77 in Production. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.79 among individual contributors, 0.75 among engagement managers, and 0.68 among portfolio directors. Among individual contributors alone, 65.06% (203 of 312) of those who rated it 1 to 3 had a score of 3 or below.

Thirty-eight respondents on 32 engagements wrote about the vendor's readiness to take accountability, one answer each, and 29 of the answers were negative, 9 from canceled and 20 from active engagements. Of the 29, 11 (37.93%) described one of the 26 forms of a control bypassed or not working, most often unclear liability or regulatory accountability (7) and problems raised with no response (6). The answers described leadership that fell silent after an error and responsibility that came to rest on the individual contributor on shift.

> *“Playbook for taking a use case into production does exist, and it is better than what we had two years ago; training on payments and card scheme rules is genuinely useful. Where I could not defend them is accountability. When model produced a summary that missed an indemnity cap and client raised it, our leadership went quiet for the better part of two weeks. That silence did more damage than the error.”*
>
> **ENGAGEMENT MANAGER**
>
> KPO, Regulatory, audit, tax and legal

> *“Two of us have covered nights since day one because tier 2 doesnt pick up. And when the agent closed out a real intrusion alert as noise, nobody upstairs put their name on it. That falls on the analyst who was on shift. Thats me some weeks. I dont sleep great”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

Part 2 (Vendor Capability, section 7.1) reports the association of the vendor's readiness to take accountability with the four study measures.

> *"training was two recorded sessions and a quiz, so my team keep our own note file, the composite rules and the checks that catch bad output are all in it. i wrote most of it. accordingly nobody ask the training team for anything now."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Trade, settlement and reconciliation

Technical Controls and Governance Across the Engagement

Technical safeguards were associated with the broader control environment, even though their low-rating share was smaller than in the management categories. The eight technology indicators had 2,052 low ratings among 6,362 ratings. The governance, oversight, and control score relates those assessments to controls elsewhere in the engagement. The main adjusted model did not detect a separate positive cross-category relationship, with sensitivity to the composition of the comparison measure.

Each technology indicator correlated with governance in the other four categories at r = 0.77–0.81 (n = 150). Under the main adjustment, partial r ranged from −0.09 to 0.08 and none met nominal $p < 0.05$. Item-to-score correlations were 0.65–0.70 in Build-Out and 0.78–0.82 in Production. This cross-sectional ordering is descriptive and does not establish a stronger relationship once the same agents enter live operation.

A distressed technology rating was rarer than a distressed management rating and associated with a larger conditional low-score share. Between 251 and 262 respondents rated a technology control 1 to 3, against 377 to 407 for each co-management indicator, and 74.90% to 78.54% of them had a score of 3 or below. Ratings of 4 to 7 still carried a score of 3 or below for 13.21% to 16.42% of respondents, against 2.65% to 5.30% for the co-management indicators. Table 3.11 sets out the eight indicators against these measures.

Table 3.11. The eight AI technology indicators against reported governance weakness across the engagement

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Resistance to Hostile Embedded Instructions	205 of 261 (78.54%)	0.86 (0.81 to 0.90)	0.81
Safe Handling of Sensitive Data	194 of 259 (74.90%)	0.84 (0.79 to 0.88)	0.79
Consistency of Intended and Authorized Agent Action	192 of 253 (75.89%)	0.83 (0.77 to 0.87)	0.78
Reliability of Built-In Guardrails in Production	193 of 252 (76.59%)	0.85 (0.79 to 0.89)	0.80
Proof of Correct Behavior Before Work Goes Live	195 of 259 (75.29%)	0.85 (0.80 to 0.89)	0.80

Table 3.11. The eight AI technology indicators against reported governance weakness across the engagement (continued)

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Accuracy of the Tools' Own Success and Failure Reports	189 of 251 (75.30%)	0.84 (0.79 to 0.88)	0.79
Explanation and Evidence of Outputs at Audit Depth	197 of 262 (75.19%)	0.86 (0.81 to 0.89)	0.81
Quality of Logs, Traces, and Usage Telemetry	191 of 255 (74.90%)	0.82 (0.76 to 0.87)	0.77

Indicator	Adjusted partial r (p)	Negative written answers
Resistance to Hostile Embedded Instructions	0.08 (p = 0.310)	9 of 9
Safe Handling of Sensitive Data	−0.09 (p = 0.295)	22 of 28
Consistency of Intended and Authorized Agent Action	−0.03 (p = 0.745)	48 of 64
Reliability of Built-In Guardrails in Production	0.06 (p = 0.488)	7 of 17
Proof of Correct Behavior Before Work Goes Live	0.05 (p = 0.527)	26 of 38
Accuracy of the Tools' Own Success and Failure Reports	−0.04 (p = 0.623)	63 of 73
Explanation and Evidence of Outputs at Audit Depth	0.03 (p = 0.739)	145 of 163
Quality of Logs, Traces, and Usage Telemetry	−0.03 (p = 0.702)	82 of 101

Note: Rating-band and item-to-score results use the score excluding the tested indicator. Cross-category results exclude its entire category and use 150 engagements. Partial r accounts for the mean of the 53 indicators outside the score. Adjusted-correlation p values are nominal. The ordinary respondent-level band tests do not account for clustering within engagements, including where Monte Carlo calculation handles sparse cells. Their fractions remain descriptive. Zero observations are not proof of no risk.

> *“In April we took a batch of eighty investigations to the bank's financial crime committee and the tool's own run summary said all eighty had completed cleanly. Six had not, they had stopped part way and reported success. Nothing came of it because our own reviewers caught it first, but i had to explain to their head of compliance why a machine's account of its own work is not evidence it did the work. That conversation set the programme back a month.”*
>
> **PORTFOLIO DIRECTOR**
>
> BPO, Financial crime and client onboarding

1 Data Isolation, Sensitive Data, and Adversarial Input

Resistance to hostile input and safe handling of sensitive data connect technical boundaries with governance of agent action. Both were associated with the wider score, while the main adjusted estimates were not clearly different from zero. These results concern reported tool properties. They do not independently establish security performance or remove the need to explain how the controls operated in the engagement.

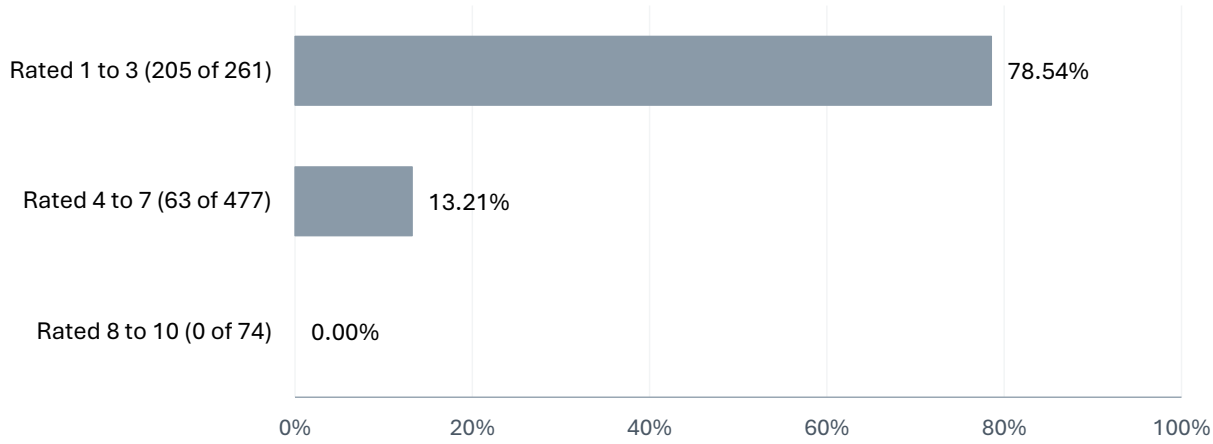
1.1 Resistance to Hostile Embedded Instructions

Indicator C167 · Rate the tools' resistance to hostile instructions embedded in the content they read and process · Full statistical analysis ›

Weak resistance to hostile embedded instructions was associated with low scores. The control concerns instructions placed inside content an agent reads, which may divert behavior from the authorized task. That mechanism connects a technical property with the governance boundary over agent action. The study records respondents' assessments of resistance, rather than results from an independent security test, and the adjusted comparison did not detect a clear separate association.

Figure 3.55. Resistance to Hostile Embedded Instructions: share with a score of 3 or below by indicator rating.

Weak resistance to hostile embedded instructions was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 261 respondents who rated resistance 1 to 3, 205 had a score of 3 or below. Among the 477 who rated it 4 to 7, 63 (13.21%) did, the lowest such share of the eight, and none of the 74 who rated it 8 to 10 did (chi-square = 365.79, $p < 0.001$, Cramér's V = 0.67, the largest conditional share of the 47 governance indicators).

Resistance was associated with the score computed without it at $r = 0.86$ (95% CI 0.81 to 0.90, $n = 150$ engagements) and with governance in the other four categories at $r = 0.81$, the joint highest of the eight. After adjustment for the mean of the other 53 indicators, the partial r was 0.08 and was not significant ($p = 0.310$). Respondents who rated it distressed recorded a score of 2.29 ($n = 261$), against 4.79 ($n = 551$) among the rest (cluster-robust $p < 0.001$).

The association ranged across the active-stage groups from 0.69 in Build-Out to 0.74 in Transition/Cut-Over and 0.81 in Production, and held within the canceled engagements at $r = 0.72$, the strongest of the eight. After Holm adjustment, it was significant in 84 of the 87 segment levels tested. Within role it stood at 0.81 among individual contributors, 0.83 among engagement managers, and 0.77 among portfolio directors. Among individual contributors alone, 80.18% (174 of 217) of those who rated it 1 to 3 had a score of 3 or below.

Nine respondents on 9 engagements wrote about hostile instructions, 8 of them on active engagements, and all nine answers were negative. None described one of the 26 forms of a control bypassed or not working, and each described the technical control itself failing. Respondents reported instructions embedded in a remittance narrative, a scanned letter, the footnotes of a circular, and the text of a ticket that the tools obeyed, in one case recording the case as closed with high confidence.

> *"Last month a circular we fed in had instruction lines in the footnotes, tool followed the same. Logs are good otherwise, I could see exactly what it ran"*
>
> **INDIVIDUAL CONTRIBUTOR**
> KPO, Regulatory, audit, tax and legal

> *"need the Tool to read runbook from one locked source... it pick up any text in ticket and follow it."*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Technology engineering and operations

Part 2 (AI Technology, section 5.3) reports the association of resistance with the four study measures.

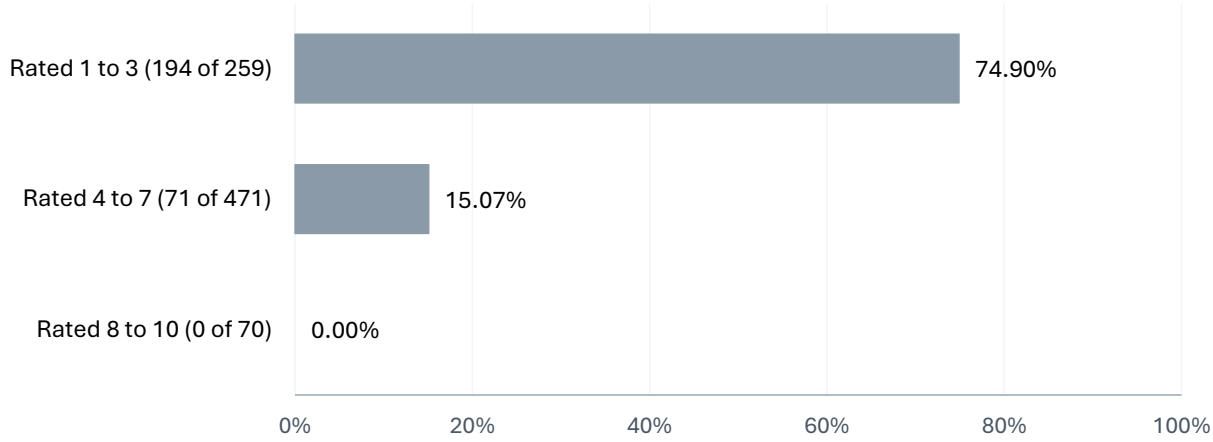
1.2 Safe Handling of Sensitive Data

Indicator C169 · Rate how safely the tools handled sensitive data, with no exposure through outputs, logs, or connected systems · Full statistical analysis ›

Low ratings for safe handling of sensitive data were associated with low scores. The indicator concerns exposure through outputs, logs, or connected systems, bringing the evidence generated by agent work within the control boundary. Its associations vary with the score and adjustment used. The nonsignificant estimates do not establish either a distinct effect or the absence of an operational data-protection problem.

Figure 3.56. Safe Handling of Sensitive Data: share with a score of 3 or below by indicator rating.

Low ratings for safe handling of sensitive data were associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

For the score excluding the sensitive-data item, the adjusted estimate was partial r = 0.11 (p = 0.179). Against governance in the other four categories, its partial r after that adjustment was −0.09 (p = 0.295), the lowest of the eight. Scores were 3 or below for 194 of 259 (74.90%) at indicator ratings 1–3. The corresponding counts were 71 of 471 (15.07%) at indicator ratings 4–7 and 0 of 70 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 308.02, p < 0.001, Cramér's V = 0.62.

The indicator was associated with the score computed without it at r = 0.84 (95% CI 0.79 to 0.88, n = 150 engagements) and with governance in the other four categories at r = 0.79. Respondents who rated it distressed recorded a score of 2.35 (n = 259), against 4.76 (n = 541) among the rest (cluster-robust p < 0.001).

The association ranged across the active-stage groups from 0.68 in Build-Out to 0.82 in Production and held within the canceled engagements at r = 0.70. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.81 among individual contributors, 0.84 among engagement managers, and 0.73 among portfolio directors. Among individual contributors alone, 75.69% (165 of 218) of those who rated it 1 to 3 had a score of 3 or below.

Twenty-eight respondents on 27 engagements wrote about sensitive data, and 22 of the answers were negative, 13 from canceled and 9 from active engagements. Three of the 22 described one of the 26 forms of a control bypassed or not working, two of them problems raised with no response. Respondents described account numbers and salary lines written in full into logs that much of the delivery team could open, and records from one account file that appeared in a summary of another.

> *“the tool put client salary lines into a log i could read... that should not be possible.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> KPO, Corporate and back-office functions

> *“it needs to stop pulling account numbers into the chat log for a start. and nobody should be letting it touch a live settlement without me reading it first”*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Trade, settlement and reconciliation

Part 2 (AI Technology, section 5.2) reports the association of sensitive data handling with the four study measures.

2 Authorized Action and Guardrails

Authorized action and production guardrails were associated with the wider score. The first concerns the scope of delegated action, while the second concerns whether restrictions held during use. Their adjusted estimates did not establish distinct relationships under the main comparison mean. Sensitivity of the authorized-action result illustrates why a conditional null should not become a conclusion that technical controls have no role of their own.

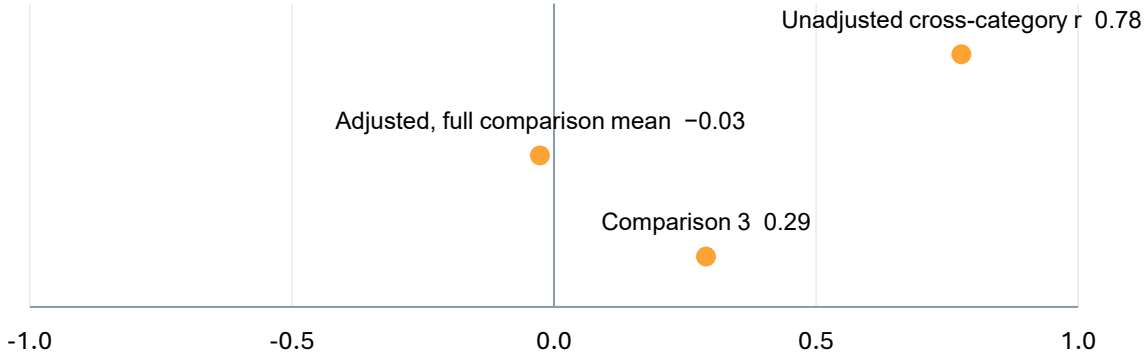
2.1 Consistency of Intended and Authorized Agent Action

Indicator C168 · Rate how consistently the agents took only intended and authorized actions when operating connected tools and systems · Full statistical analysis ›

Inconsistent restriction of agent activity to intended and authorized actions was associated with low scores. The control function is the boundary between delegated work and action outside the delegation. The adjusted cross-category estimate was near zero under the main comparison mean, but positive in the category-removal sensitivity check. This model dependence limits a claim that authorization has no relationship of its own, without establishing a causal effect.

Figure 3.57. Consistency of Intended and Authorized Agent Action: observed association and adjustment sensitivity.

Inconsistent restriction of agent activity to intended and authorized actions was associated with low scores.



Comparison	Estimate and analysis base
Unadjusted cross-category r	r = 0.78 150 engagements
Adjusted, full comparison mean	r = −0.03 150 engagements
Adjusted, excluding other AI technology items	r = 0.29 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

Correlations use 150 engagements. The alternative mean removes the stated category items. Estimates are descriptive model comparisons, with nominal inference in the text.

Of the 253 respondents who rated authorized action 1 to 3, 192 had a score of 3 or below. Among the 474 who rated it 4 to 7, 69 (14.56%) did, and none of the 62 who rated it 8 to 10 did (chi-square = 313.56, $p < 0.001$, Cramér's V = 0.63).

Authorized action was associated with the score computed without it at r = 0.83 (95% CI 0.77 to 0.87, n = 150 engagements) and with governance in the other four categories at r = 0.78. After adjustment for the mean of the other 53 indicators, the partial r was −0.03 (p = 0.745). Respondents who rated it distressed recorded a score of 2.31 (n = 253), against 4.78 (n = 536) among the rest (cluster-robust p < 0.001). Removing the other 12 AI technology items from the comparison mean changed partial r to 0.29 (nominal p < 0.001), illustrating model sensitivity.

The descriptive stage correlations ranged from 0.65 in Build-Out to 0.71 in Transition/Cut-Over and 0.78 in Production and held within the canceled engagements at r = 0.71. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.81 among individual contributors, 0.75 among engagement managers, and 0.70 among portfolio directors. Among individual contributors alone, 78.20% (165 of 211) of those who rated it 1 to 3 had a score of 3 or below.

Authorized action tracked the manual alteration of AI outputs. Of the respondents who rated it 1 to 3, 66.39% (158 of 238) altered AI outputs, against 18.76% (88 of 469) at ratings of 4 to 7 and 8.06% (5 of 62) at ratings of 8 to 10 (chi-square p < 0.001).

Authorized action drew 64 written answers from 62 respondents on 50 engagements, and 48 were negative, 19 from canceled and 29 from active engagements. Of the 48, 7 (14.58%) described one of the 26 forms of a control bypassed or not working, most often problems raised with no response (3). Respondents described agents that pulled records outside the batch they were given, touched queues and accounts beyond their scope, and in one account did so through a loose instruction rather than a failed control.

> *“it gives same answer more often than before, that part improved. but it still touches queues it has no business touching, and nothing is logged.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Technology engineering and operations

Part 2 (AI Technology, section 6.1) reports the association of authorized action with the four study measures.

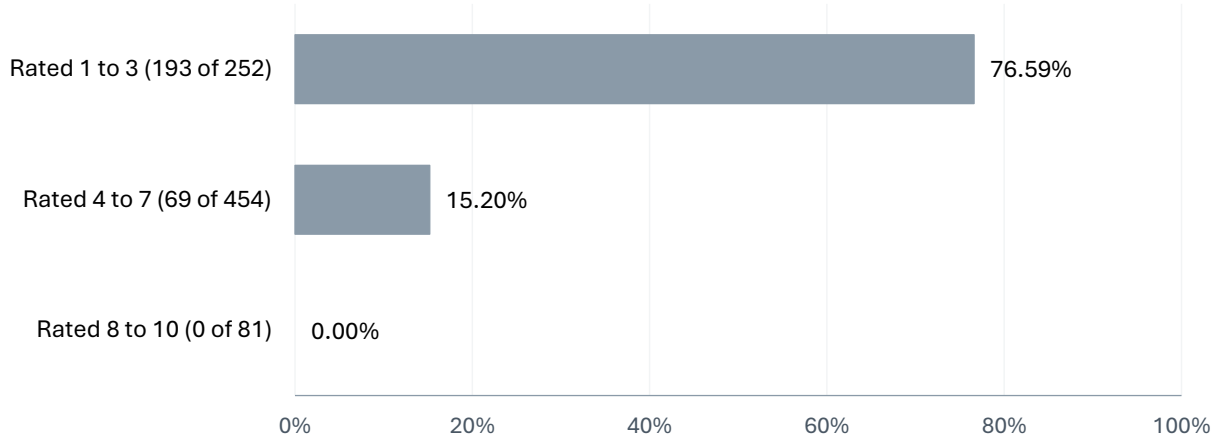
2.2 Reliability of Built-In Guardrails in Production

Indicator C170 · Rate how reliably the tools' built-in guardrails and safety controls held during production use · Full statistical analysis ›

Less reliable built-in guardrails were associated with low scores. The indicator concerns whether safeguards held during production use, linking designed restrictions with reported operation. Its findings should be distinguished from the presence of a documented safety policy and from authorized-action ratings. The main adjusted comparison did not detect a clear separate association, and the relatively less negative written answers do not independently verify guardrail effectiveness.

Figure 3.58. Reliability of Built-In Guardrails in Production: share with a score of 3 or below by indicator rating.

Less reliable built-in guardrails were associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 252 respondents who rated guardrails 1 to 3, 193 had a score of 3 or below, the second-highest share of the eight. Among the 454 who rated them 4 to 7, 69 (15.20%) did, and none of the 81 who rated them 8 to 10 did (chi-square = 320.05, $p < 0.001$, Cramér's V = 0.64).

Guardrails were associated with the score computed without them at r = 0.85 (95% CI 0.79 to 0.89, n = 150 engagements) and with governance in the other four categories at r = 0.80. After adjustment for the mean of the other 53 indicators, the partial r was 0.06 ($p = 0.488$). Respondents who rated them distressed recorded a score of 2.35 (n = 252), against 4.78 (n = 535) among the rest (cluster-robust $p < 0.001$).

The association ranged across the active-stage groups from 0.66 in Build-Out to 0.80 in Production and held within the canceled engagements at r = 0.71. After Holm adjustment, it

was significant in 85 of the 87 segment levels tested. Within role it stood at 0.82 among individual contributors and 0.79 among engagement managers and portfolio directors. Among individual contributors alone, 78.74% (163 of 207) of those who rated them 1 to 3 had a score of 3 or below.

Seventeen respondents on 14 engagements wrote about guardrails, and 7 of the answers were negative, all from active engagements. One of the 7 linked the failure to regulatory or compliance exposure. Respondents reported guardrails that now held while the tool lost instructions within a flow, and guardrails that stayed the weak point of a tool whose matching had improved.

> *“Guardrails hold, thats better than it was. The fnol flow still loses what I told it two steps back.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Client servicing and communications

> *“the matching itself got much better over the six months, it runs a full statement end to end now and when it says it cleared the break it really did clear it. the guardrails are the weak part still, it will happily action something on a client cash account that i would want a second pair of eyes on”*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Trade, settlement and reconciliation

Part 2 (AI Technology, section 6.2) reports the association of guardrails with the four study measures.

“report job runs, fails halfway, runs again from start. Same job three times in one morning is normal here, furthermore nobody can tell me why because there is no trace, only a status of failed. Logs from Platform show start time and end time and nothing between. When it does finish, numbers are mostly right, ill give it that. But i cant prove anything to my lead, i can only re-run it and show him the third one.”

INDIVIDUAL CONTRIBUTOR

KPO, Corporate and back-office functions

3 Testability and Behavioral Stability

Testability connects assurance before live work with the broader score. The indicator asks whether correct behavior can be established through testing, while the later output-consistency analysis concerns variation during repeated work. Together they distinguish prospective assurance from observed stability. A favorable rating remains a report about capability, not evidence that all relevant behavior was tested successfully.

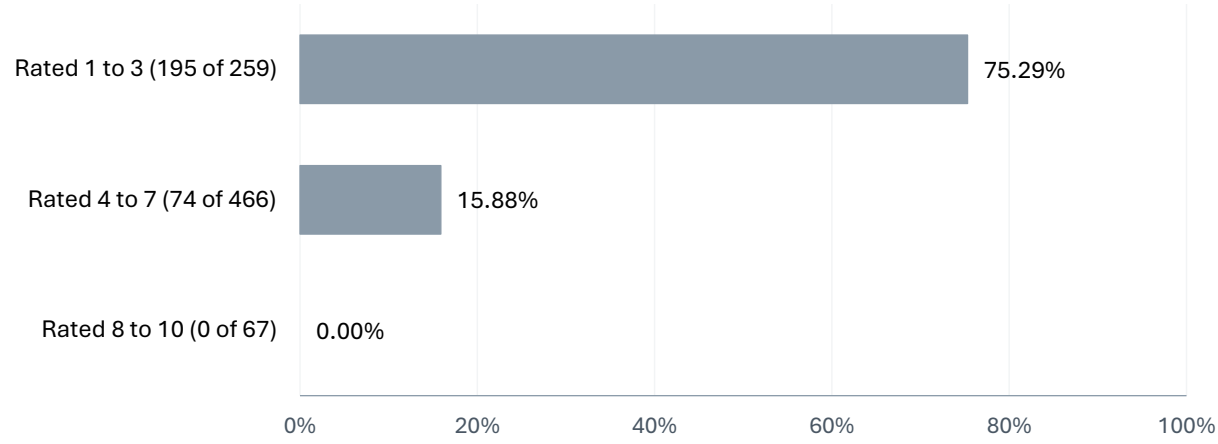
3.1 Proof of Correct Behavior Before Work Goes Live

Indicator C174 · Rate how well the tools supported reliable testing, meaning correct behavior could be proven before work went live · Full statistical analysis ›

Limited support for testing correct behavior before live operation was associated with low scores. This connects the ability to examine agent work before deployment with the wider reported control environment. A favorable rating describes perceived testability, not proof that all relevant behavior was tested or that correctness was established. The adjusted result qualifies the association without resolving that operational assurance question.

Figure 3.59. Proof of Correct Behavior Before Work Goes Live: share with a score of 3 or below by indicator rating.

Limited support for testing correct behavior before live operation was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 259 respondents who rated testability 1 to 3, 195 had a score of 3 or below. Among the 466 who rated it 4 to 7, 74 (15.88%) did, and none of the 67 who rated it 8 to 10 did (chi-square = 299.62, $p < 0.001$, Cramér's V = 0.62).

Testability was associated with the score computed without it at $r = 0.85$ (95% CI 0.80 to 0.89, n = 150 engagements) and with governance in the other four categories at $r = 0.80$. After adjustment for the mean of the other 53 indicators, the partial r was 0.05 ($p = 0.527$). Respondents who rated it distressed recorded a score of 2.34 (n = 259), against 4.75 (n = 533) among the rest (cluster-robust $p < 0.001$).

The association ranged across the active-stage groups from 0.67 in Build-Out to 0.73 in Transition/Cut-Over and 0.80 in Production and held within the canceled engagements at $r = 0.71$. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.81 among individual contributors and engagement managers and 0.74 among portfolio directors. Among individual contributors alone, 76.82% (169 of 220) of those who rated it 1 to 3 had a score of 3 or below.

Thirty-eight respondents on 33 engagements wrote about testability, and 26 of the answers were negative, 12 from canceled and 14 from active engagements. Three of the 26 described one of the 26 forms of a control bypassed or not working. Respondents described the absence of any test environment, validation performed on live cases, and a reviewer's own reading as the only proof that an agent's work was correct.

> *"Whoever reads this, the main thing is we cannot prove anything works before it goes live, there is no real test here, only my eyes"*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Regulatory, audit, tax and legal

> *"there was no way to test properly before going live. we found the misses later from the false positive report. i would ask you to note that is the main thing here"*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Financial crime and client onboarding

Part 2 (AI Technology, section 7.1) reports the association of testability with the four study measures.

4 Observability of Agent Work

Observability connects agent work with the evidence available to oversee it. Tool reports, output explanations, and logs were associated with the wider score, while tool-reported success also correlated with the accuracy of reporting supplied to the financial institution. These measures concern different parts of the evidence chain. Their alignment does not independently verify the completeness or truth of the underlying record.

> *"Main thing is it report success when it has not done the thing. Pipeline log say green, the artefact is not in the repo... that is the one that cost us most"*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Technology engineering and operations

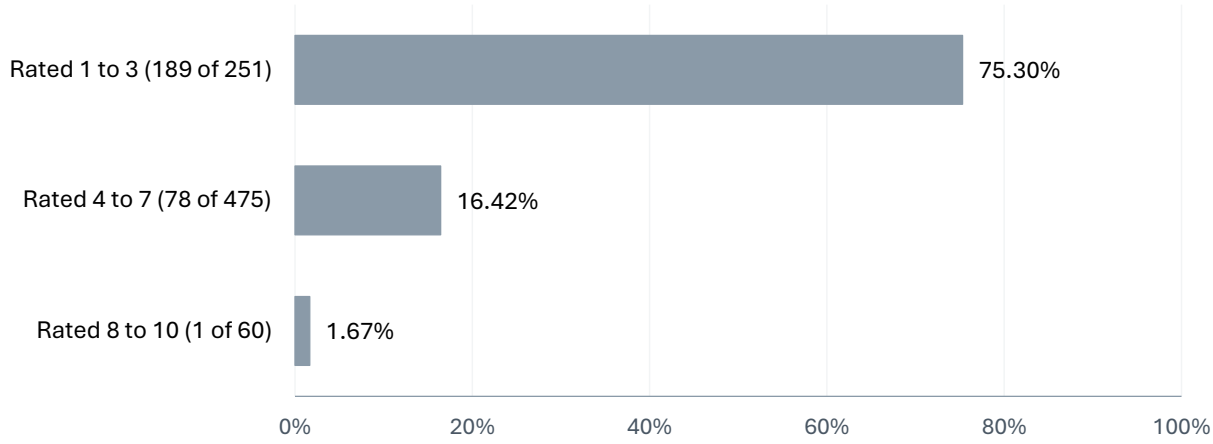
4.1 Accuracy of the Tools' Own Success and Failure Reports

Indicator C172 · Rate how accurately the tools' own reports of success or failure matched the actual outcomes of their work · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 6.8 · Full statistical analysis ›

Inaccurate tool reports of success and failure were associated with low scores. This creates a governance concern because oversight depends on evidence that corresponds to the work performed. A system can report successful completion while producing an incorrect or absent result. The relationship with reporting provided to the financial institution links this technical evidence problem with the information used to oversee delivery.

Figure 3.60. Accuracy of the Tools' Own Success and Failure Reports: share with a score of 3 or below by indicator rating.

Inaccurate tool reports of success and failure were associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Among the 251 respondents rating tool-report accuracy at 1–3, 189 had a score of 3 or below (75.30%). The corresponding counts were 78 of 475 at ratings 4–7 (16.42%) and one of 60 at ratings 8–10. The ordinary band test was chi-square = 283.75 ($p < 0.001$, Cramér's V = 0.60), subject to the stated clustering limitation. The reporting-accuracy correlation of 0.68 addresses a separate relationship.

The tools' reports were associated with the score computed without them at $r = 0.84$ (95% CI 0.79 to 0.88, n = 150 engagements) and with governance in the other four categories at $r = 0.79$. After adjustment for the mean of the other 53 indicators, the partial r was −0.04 ($p = 0.623$). Respondents who rated them distressed recorded a score of 2.30 (n = 251), against 4.74 (n = 535) among the rest (cluster-robust $p < 0.001$).

The association ranged across the active-stage groups from 0.67 in Build-Out to 0.81 in Production and held within the canceled engagements at r = 0.69. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.79 among individual contributors, 0.81 among engagement managers, and 0.78 among portfolio directors. Among individual contributors alone, 77.25% (163 of 211) of those who rated them 1 to 3 had a score of 3 or below.

The tools' own reports tracked reporting alteration. Of the respondents who rated them 1 to 3, 82.87% (208 of 251) described reporting alteration, against 46.32% (220 of 475) at ratings of 4 to 7 and 15.00% (9 of 60) at ratings of 8 to 10 (chi-square $p < 0.001$).

The indicator drew 73 written answers from 73 respondents on 58 engagements, and 63 were negative, 27 from canceled and 36 from active engagements. Of the 63, 10 (15.87%) described one of the 26 forms of a control bypassed or not working, most often change discovered only through failure (5) and problems raised with no response (4). Respondents described runs reported as complete in which items had never been processed, and teams that stopped trusting the status and counted the output themselves.

> *"The job status is the main issue. Tool reports the load as completed, then we check the target table and half the rows are not there. This happened on four occasions in my time here. Logs do not show anything useful, only that the step finished. So we stopped trusting the status and started counting rows ourselves each morning"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

> *"To be fair, by month six the behaviour had settled -- it stopped drifting week to week, and a composite run today looks much like the same run last month. What never improved is the reporting it does on itself. It would finish a batch of disclosures and mark it complete when three of the composites had not been built at all, and the logs give you nothing to work back from. We only found those by hand, on the day the pack was due."*
>
> **ENGAGEMENT MANAGER**
>
> KPO, Trade, settlement and reconciliation

Part 2 (AI Technology, section 8.1) reports the association of the tools' reports with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 6.8) their association with manual alteration.

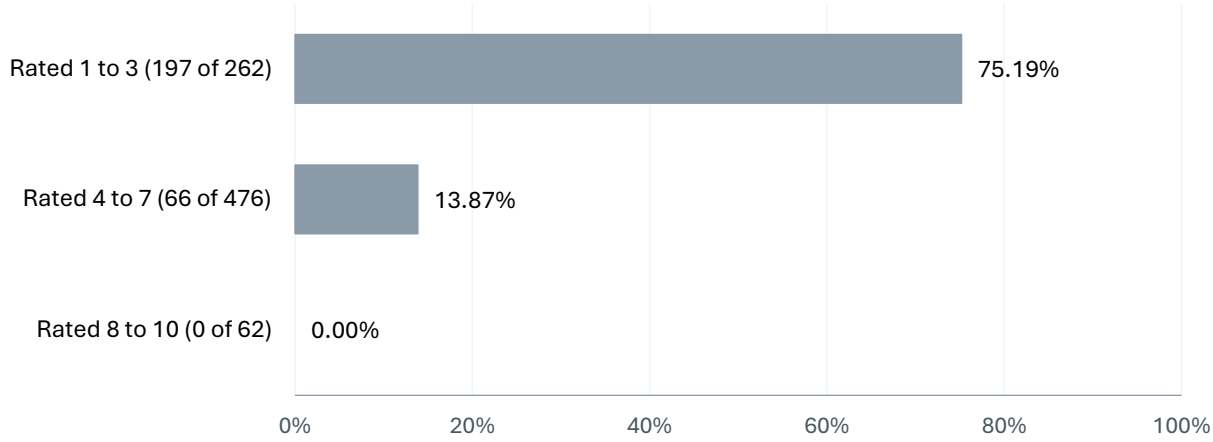
4.2 Explanation and Evidence of Outputs at Audit Depth

Indicator C173 · Rate the tools' ability to explain and evidence how each output was produced, at the depth an audit requires · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 6.5 · Full statistical analysis ›

Weak explanation and evidence of how outputs were produced were associated with low scores. The indicator concerns whether agent work can be reconstructed at the depth an audit requires. This is distinct from whether an output appears correct or a dashboard reports success. The ratings describe perceived auditability, with no independent finding that an audit standard or legal obligation was met or breached.

Figure 3.61. Explanation and Evidence of Outputs at Audit Depth: share with a score of 3 or below by indicator rating.

Weak explanation and evidence of how outputs were produced were associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 262 respondents who rated explanation 1 to 3, 197 had a score of 3 or below. Among the 476 who rated it 4 to 7, 66 (13.87%) did, and none of the 62 who rated it 8 to 10 did (chi-square = 320.91, $p < 0.001$, Cramér's V = 0.63).

Explanation was associated with the score computed without it at r = 0.86 (95% CI 0.81 to 0.89, n = 150 engagements) and with governance in the other four categories at r = 0.81. After adjustment for the mean of the other 53 indicators, the partial r was 0.03 ($p = 0.739$). Respondents who rated it distressed recorded a score of 2.34 (n = 262), against 4.79 (n = 538) among the rest (cluster-robust $p < 0.001$).

The association ranged across the active-stage groups from 0.70 in Build-Out to 0.77 in Transition/Cut-Over and 0.81 in Production and held within the canceled engagements at r =

0.71. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.82 among individual contributors, 0.80 among engagement managers, and 0.87 among portfolio directors. Among individual contributors alone, 76.82% (169 of 220) of those who rated it 1 to 3 had a score of 3 or below.

Explanation drew 163 written answers from 152 respondents on 90 engagements, and 145 were negative, 65 from canceled and 80 from active engagements. Only 9 of the 145 (6.21%) described one of the 26 forms of a control bypassed or not working, so the answers recorded a control that the tools could not supply rather than one the delivery team set aside. Respondents described compliance and audit questions answered by rebuilding the trail manually, outputs that could be corrected and not explained, and citations to clauses that did not exist.

> *“It cannot show where a number came from... i asked for the working and got a summary only. On a long custodian statement it skips whatever sits in the middle pages”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Trade, settlement and reconciliation

> *“Checked our exception log before answering. There is no rating for this, when it files document to wrong class we cannot show why afterwards. We can correct it, we cannot explain it, and for records work explaining is half the job.”*
>
> **ENGAGEMENT MANAGER**
>
> BPO, Corporate and back-office functions

Part 2 (AI Technology, section 8.2) reports the association of explanation with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 6.5) its association with manual alteration.

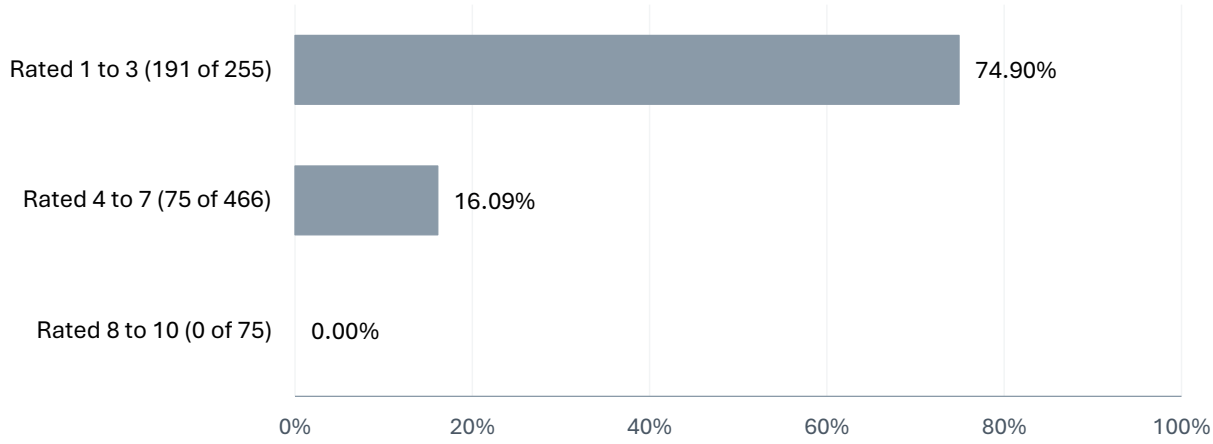
4.3 Quality of Logs, Traces, and Usage Telemetry

Indicator C176 · Rate the quality of the tools' logs, traces, and usage telemetry for detecting and diagnosing failures · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 6.6 · Full statistical analysis ›

Weak logs, traces, and usage telemetry were associated with low scores. These records provide the evidence needed to detect a problem and examine what an agent did. Their governance significance lies in observability and diagnosis, rather than the volume of records generated. The separate joint comparison with output consistency examines how two reported limitations occur together and does not establish that one causes or amplifies the other.

Figure 3.62. Quality of Logs, Traces, and Usage Telemetry: share with a score of 3 or below by indicator rating.

Weak logs, traces, and usage telemetry were associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 255 respondents who rated logs and telemetry 1 to 3, 191 had a score of 3 or below, the same share as for sensitive data handling. Among the 466 who rated them 4 to 7, 75 (16.09%) did, and none of the 75 who rated them 8 to 10 did (chi-square = 297.73, $p < 0.001$, Cramér's V = 0.61).

Logs were associated with the score computed without them at $r = 0.82$ (95% CI 0.76 to 0.87, n = 150 engagements), the lowest of the eight, and with governance in the other four categories at $r = 0.77$. After adjustment for the mean of the other 53 indicators, the partial r was −0.03 ($p = 0.702$). Respondents who rated them distressed recorded a score of 2.36 (n = 255), against 4.76 (n = 541) among the rest (cluster-robust $p < 0.001$).

The association ranged across the active-stage groups from 0.66 in Build-Out to 0.72 in Transition/Cut-Over and 0.79 in Production. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.80 among individual contributors and 0.76 among engagement managers and fell to 0.61 among portfolio directors. Among individual contributors alone, 77.93% (166 of 213) of those who rated them 1 to 3 had a score of 3 or below.

Logs and output-consistency ratings were strongly correlated across engagements ($r = 0.94$, $n = 150$). Their joint comparison, reported under The Fit Between Governance Processes and the Pace of Agent Work, uses only respondents with both numeric answers. It describes the coincidence of reported limitations and does not establish that either caused the other.

Logs drew 101 written answers from 97 respondents on 73 engagements, and 82 were negative, 47 from canceled and 35 from active engagements. Of the 82, 9 (10.98%) described one of the 26 forms of a control bypassed or not working. Respondents described logs that recorded a start and a finish and nothing between, and failures whose point of origin could not be found.

> *“Last month I gave it the recertification batch, two hundred accounts. Around ninety it forgot the rule on contractors and started passing them through... and nothing in the logs tells me where it turned”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

Part 2 (AI Technology, section 8.3) reports the association of logs with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 6.6) their association with manual alteration.

“

To be fair to the providers they do send the deprecation notices well in advance, we just had nobody free to action them. The bigger problem is half their systems only have a screen, no proper api, so the agent has to click like a person”

INDIVIDUAL CONTRIBUTOR

ITO, Technology engineering and operations

Third-Party Dependencies and the Limits of Internal Control

External dependencies were associated with the broader control environment through notice, monitoring, contractual coverage, security, and operational knowledge. The ten third-party governance indicators had 2,533 low ratings among 7,944 ratings. The governance, oversight, and control score relates these dependencies to the engagement's other controls. Written accounts of change discovered through failure add a specific detection concern, while adjusted associations remain conditional and less conclusive.

Among respondents who rated a third-party control 1 to 3, 73.41% to 80.41% had a score of 3 or below, the highest range of the five categories. Ratings of 4 to 7 still carried a score of 3 or below for 14.14% to 17.03% of respondents, and the upper end of that range was the highest of the 47 governance indicators. Each indicator was associated with governance in the other four failure categories at r = 0.75 to 0.78 (n = 150 engagements), and after adjustment for the mean of the other 53 indicators the partial r ran from −0.14 to −0.04, with none of the ten significant.

Written accounts of change discovered through failure occurred on 93 of the 150 engagements and included six of the 39 respondents above score 7. Respondent mean scores were not clearly different between those describing that pattern and those not describing it. Unannounced schema and format changes similarly showed no clearly detected mean contrast. These results preserve a specific detection concern across the observed score range, without demonstrating equal incidence. Table 3.12 reports the ten controls and two written patterns.

Table 3.12. The ten third-party ecosystem indicators and two written patterns against reported governance weakness across the engagement

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Quality of Advance Notice From Third-Party Providers	192 of 250 (76.80%)	0.85 (0.80 to 0.89)	0.77
Delivery of Change Notices to the Operators of Affected Agents	188 of 255 (73.73%)	0.83 (0.78 to 0.88)	0.75
Protection of Agents From Silent Failure	186 of 249 (74.70%)	0.85 (0.80 to 0.89)	0.78

Table 3.12. The ten third-party ecosystem indicators and two written patterns against reported governance weakness across the engagement (continued)

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Monitoring for Unannounced Third-Party Change	188 of 251 (74.90%)	0.84 (0.79 to 0.88)	0.76
Contractual Coverage of Disruptive Provider Change	199 of 256 (77.73%)	0.84 (0.79 to 0.88)	0.76
Security Posture of Third-Party Tools and Sub-Processors	192 of 251 (76.49%)	0.84 (0.78 to 0.88)	0.76
Management of Dependence on a Small Number of Model and Cloud Providers	199 of 257 (77.43%)	0.84 (0.79 to 0.88)	0.76
Replacement of Informal Human Fixes With Explicit Checks	201 of 267 (75.28%)	0.85 (0.80 to 0.89)	0.77
Mapping of Third- and Fourth-Party Dependencies	185 of 252 (73.41%)	0.84 (0.79 to 0.88)	0.76
Management of Dependence on Gated External Rails	197 of 245 (80.41%)	0.84 (0.78 to 0.88)	0.76
Change discovered only through failure	3.82 against 4.03 (describing against the rest)	−0.17 (engagement share)	
Schema or format changed without notice	3.70 against 4.01 (describing against the rest)	−0.16 (engagement share)	

Indicator	Adjusted partial r (p)	Negative written answers
Quality of Advance Notice From Third-Party Providers	−0.09 (p = 0.277)	108 of 160
Delivery of Change Notices to the Operators of Affected Agents	−0.11 (p = 0.168)	42 of 56
Protection of Agents From Silent Failure	−0.06 (p = 0.457)	100 of 136
Monitoring for Unannounced Third-Party Change	−0.12 (p = 0.136)	30 of 46

Table 3.12. The ten third-party ecosystem indicators and two written patterns against reported governance weakness across the engagement (continued)

Indicator	A score of 3 or below among respondents rating it 1 to 3	r with score excluding item (95% CI)	r with the other four categories
Contractual Coverage of Disruptive Provider Change		−0.07 (p = 0.406)	14 of 21
Security Posture of Third-Party Tools and Sub-Processors		−0.14 (p = 0.083)	4 of 9
Management of Dependence on a Small Number of Model and Cloud Providers		−0.04 (p = 0.646)	19 of 26
Replacement of Informal Human Fixes With Explicit Checks		−0.11 (p = 0.183)	7 of 9
Mapping of Third- and Fourth-Party Dependencies		−0.13 (p = 0.105)	35 of 44
Management of Dependence on Gated External Rails		−0.08 (p = 0.333)	13 of 15
Change discovered only through failure		−0.04 (p = 0.657)	157 answers
Schema or format changed without notice		−0.06 (p = 0.478)	61 answers

Note: Rated controls use the score excluding the tested indicator for rating-band and item-to-score comparisons. Cross-category comparisons exclude the entire third-party category and use 150 engagements. Partial r accounts for the mean of the other 53 indicators, with nominal p values. Written-pattern contrasts compare respondents describing the pattern with those not describing it, rather than a verified nonincident group. Ordinary band tests do not account for engagement clustering.

> *“On my last job the network sent a bulletin six weeks before a mandate change and you planned around it. Here we found out from a failed auth response. Same networks. The difference is nobody set up the notifications.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Payments, cards and disputes

“

When our model provider retired a version this spring, we had thirty days' notice and the reclassification of the client's scanned document types had to be revalidated inside it. Nothing broke, but we have one provider and no second path if that window ever gets shorter.”

PORTFOLIO DIRECTOR

BPO, Corporate and back-office functions

1 Notice of Provider Change

Provider notice and delivery of that notice to operators were associated with the wider score. They are separate control steps: a provider may communicate a change without the people running the affected agents receiving or acting on it. The written accounts place part of the notice problem within the engagement's own arrangements. This distinction gives operational meaning to the ratings beyond a simple measure of provider communication quality.

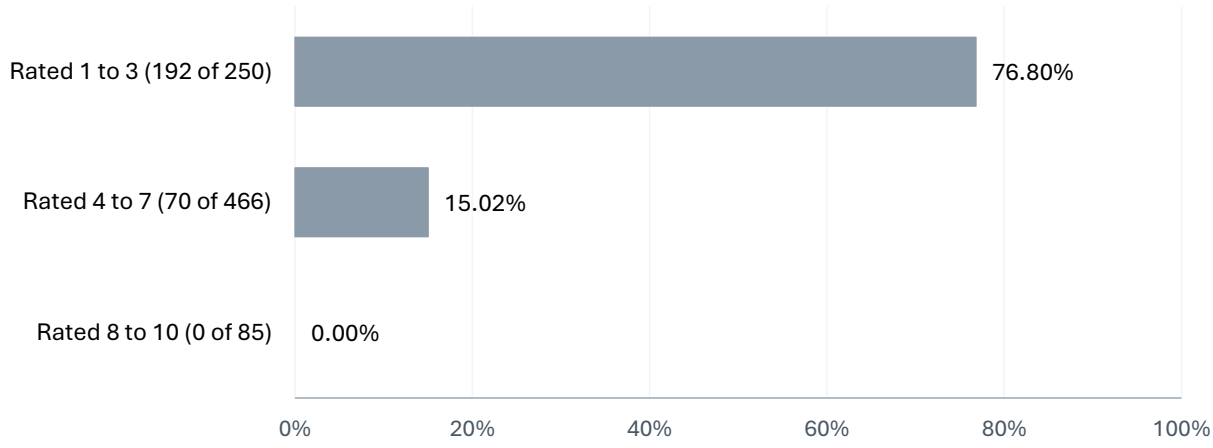
1.1 Quality of Advance Notice From Third-Party Providers

Indicator C181 · Rate the quality of advance notice third-party providers give before changes that can affect the agents on this engagement · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 7.4 · Full statistical analysis ›

Poor advance notice of provider change was associated with low scores. This places an external dependency inside the governance analysis: the financial institution and vendor need time to assess a change that can affect their agents. Notice quality is distinct from whether the message reaches operators and from whether the engagement acts on it. The indicator measures reported notice rather than every provider change event.

Figure 3.63. Quality of Advance Notice From Third-Party Providers: share with a score of 3 or below by indicator rating.

Poor advance notice of provider change was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 250 respondents who rated advance notice 1 to 3, 192 had a score of 3 or below. Among the 466 who rated it 4 to 7, 70 (15.02%) did, and none of the 85 who rated it 8 to 10 did (chi-square = 328.36, $p < 0.001$, Cramér's V = 0.64).

Advance notice was associated with the score computed without it at $r = 0.85$ (95% CI 0.80 to 0.89, n = 150 engagements) and with governance in the other four categories at $r = 0.77$. After adjustment for the mean of the other 53 indicators, the partial r was −0.09 ($p = 0.277$). Respondents who rated it distressed recorded a score of 2.33 (n = 250), against 4.76 (n = 551) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at $r = 0.72$ and across the active stages at 0.73 in Build-Out, 0.69 in Transition/Cut-Over, and 0.78 in Production. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.82 among individual contributors, 0.77 among engagement managers, and 0.76 among portfolio directors. Among individual contributors alone, 77.93% (166 of 213) of those who rated it 1 to 3 had a score of 3 or below.

Advance notice drew 160 written answers from 159 respondents on 96 engagements, and 108 were negative, 53 from canceled and 55 from active engagements. Of the 108, 83 (76.85%) described one of the 26 forms of a control bypassed or not working, most often change discovered only through failure (60) and schema or format changed without notice (41). Among the 52 answers that were not negative, several credited the providers with notice and placed the failure on the engagement's own side, where no person was assigned to read the notices that arrived.

> *“to be fair the screening provider does send notice when their list format change. nobody on our side is given the job of reading it, and so we find out when a batch fail. that part is not on them”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Financial crime and client onboarding

Part 2 (Third-Party Ecosystem, section 1.1) reports the association of advance notice with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 7.4) its association with manual alteration.

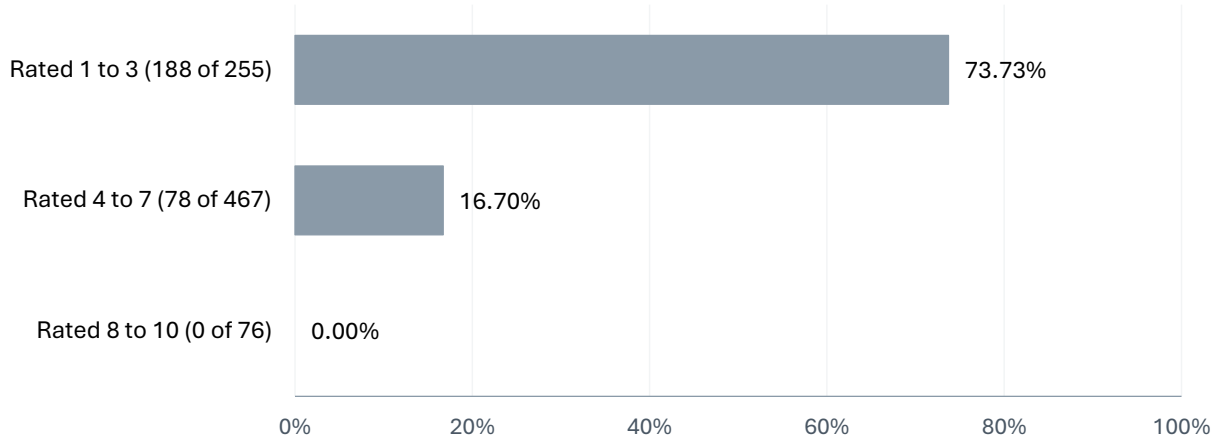
1.2 Delivery of Change Notices to the Operators of Affected Agents

Indicator C182 · Rate how reliably word of an upcoming third-party change reaches the people who operate the affected agents in time to act · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 7.8 · Full statistical analysis ›

Failure to deliver change notices to affected operators was associated with low scores. This identifies a control step inside the engagement even when a third-party provider gives notice. Information must reach people who can assess the effect on agent work before the change takes effect. The finding separates the existence of communication from its timely operational use.

Figure 3.64. Delivery of Change Notices to the Operators of Affected Agents: share with a score of 3 or below by indicator rating.

Failure to deliver change notices to affected operators was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Within canceled engagements, the correlation was 0.65, below its value in each active stage, at 0.78 in Build-Out, 0.70 in Transition/Cut-Over, and 0.74 in Production. After Holm adjustment, it was significant in 86 of the 87 segment levels tested.

Scores were 3 or below for 188 of 255 (73.73%) at indicator ratings 1–3. The corresponding counts were 78 of 467 (16.70%) at indicator ratings 4–7 and 0 of 76 (0.00%) at indicator ratings 8–10. The reported band results were chi-square = 283.34, $p < 0.001$, Cramér's V = 0.60. The indicator was associated with the score computed without it at $r = 0.83$ (95% CI 0.78 to 0.88, $n = 150$ engagements) and with governance in the other four categories at $r = 0.75$, the lowest of the ten. After adjustment for the mean of the other 53 indicators, the partial r was −0.11 ($p = 0.168$).

Within role the association stood at 0.79 among individual contributors, 0.80 among engagement managers, and 0.84 among portfolio directors. Among individual contributors alone, 76.44% (159 of 208) of those who rated it 1 to 3 had a score of 3 or below. Respondents who rated it distressed recorded a score of 2.37 (n = 255), against 4.72 (n = 543) among the rest (cluster-robust $p < 0.001$).

The indicator drew 56 written answers from 55 respondents on 47 engagements, and 42 were negative, 22 from canceled and 20 from active engagements. Of the 42, 24 (57.14%) described one of the 26 forms of a control bypassed or not working, most often change discovered only through failure (18). Respondents described notices that the providers sent and that never reached the team operating the agents, and a delivery team that learned of a change from its own logs.

> *"Last release the market data vendor changed a price field type and our pricing module threw errors all morning. notice was sent apparently, it just never reached my team, we found it from the logs ourself"*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Technology engineering and operations

> *"The API people are not the villains here. They publish their changes, it is our own side that never passes it down to the floor... not their fault."*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Client servicing and communications

Part 2 (Third-Party Ecosystem, section 1.2) reports the association of the delivery of notices with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 7.8) its association with manual alteration.

2 Detection of Unannounced Change

Protection and monitoring were associated with the wider score, while written descriptions identified changes discovered only after failure. This brings detection into the governance question. Advance notice cannot account for changes that arrive without warning, so the engagement's ability to identify disruption remains a separate responsibility. The written comparisons do not establish that late discovery was equally prevalent across score levels.

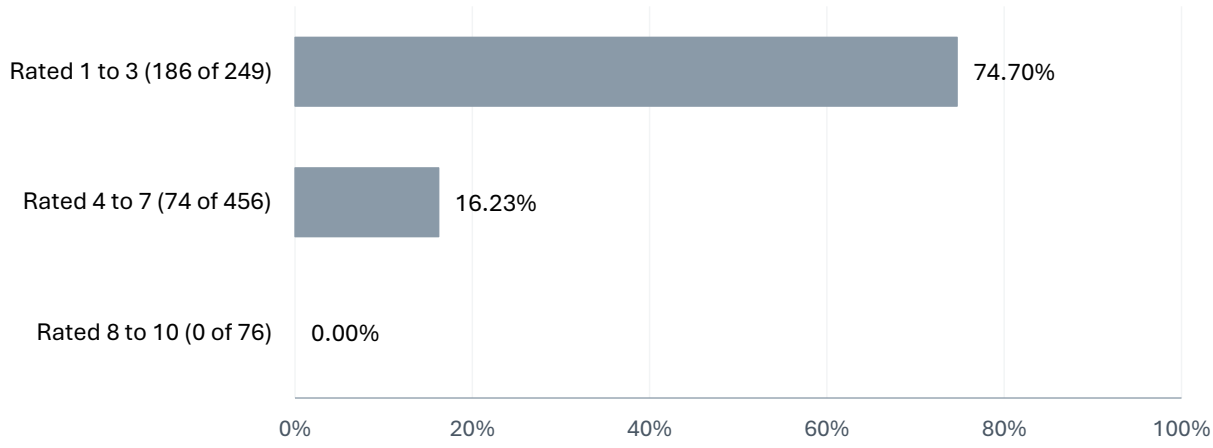
2.1 Protection of Agents From Silent Failure

Indicator C178 · Rate how well the agents on this engagement are protected from breaking silently when a third party changes something without warning · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 7.5 · Full statistical analysis ›

Weak protection against silent breakage was associated with low scores. The governance issue is whether external change can disrupt agent work without timely detection. This differs from advance notice, because protection is still needed when providers do not warn the engagement. The rating concerns perceived protection and does not count all silent failures or establish their independently observed frequency.

Figure 3.65. Protection of Agents From Silent Failure: share with a score of 3 or below by indicator rating.

Weak protection against silent breakage was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 249 respondents who rated protection 1 to 3, 186 had a score of 3 or below. Among the 456 who rated it 4 to 7, 74 (16.23%) did, and none of the 76 who rated it 8 to 10 did (chi-square = 289.95, $p < 0.001$, Cramér's V = 0.61).

Protection was associated with the score computed without it at $r = 0.85$ (95% CI 0.80 to 0.89, n = 150 engagements) and with governance in the other four categories at $r = 0.78$, the highest of the ten. After adjustment for the mean of the other 53 indicators, the partial r was −0.06 ($p = 0.457$). Respondents who rated it distressed recorded a score of 2.40 (n = 249), against 4.76 (n = 532) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at $r = 0.70$ and across the active stages at 0.74 to 0.77, and after Holm adjustment it held in 85 of the 87 segment levels tested. Within role it stood at 0.82 among individual contributors, 0.81 among engagement managers, and 0.78 among portfolio directors. Among individual contributors alone, 74.64% (156 of 209) of those who rated it 1 to 3 had a score of 3 or below.

Protection drew 136 written answers from 134 respondents on 81 engagements, and 100 were negative, 50 from canceled and 50 from active engagements. Of the 100, 74 described one of the 26 forms of a control bypassed or not working, and 66 of them described change discovered only through failure. Respondents described provider releases over a weekend that left agents processing nothing, or processing wrongly, until a person noticed on the next working day.

> *“the ticketing tool pushed a new api version on a saturday. our agent kept closing change tickets that were never raised, and nobody noticed till monday cutover. the old release guy used to catch that by eye”*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Technology engineering and operations

> *“The OCR vendor pushed a version change over a weekend and no one told us... we found it Monday when half the scans came back with blank fields.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Corporate and back-office functions

Part 2 (Third-Party Ecosystem, section 2.1) reports the association of protection with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 7.5) its association with manual alteration.

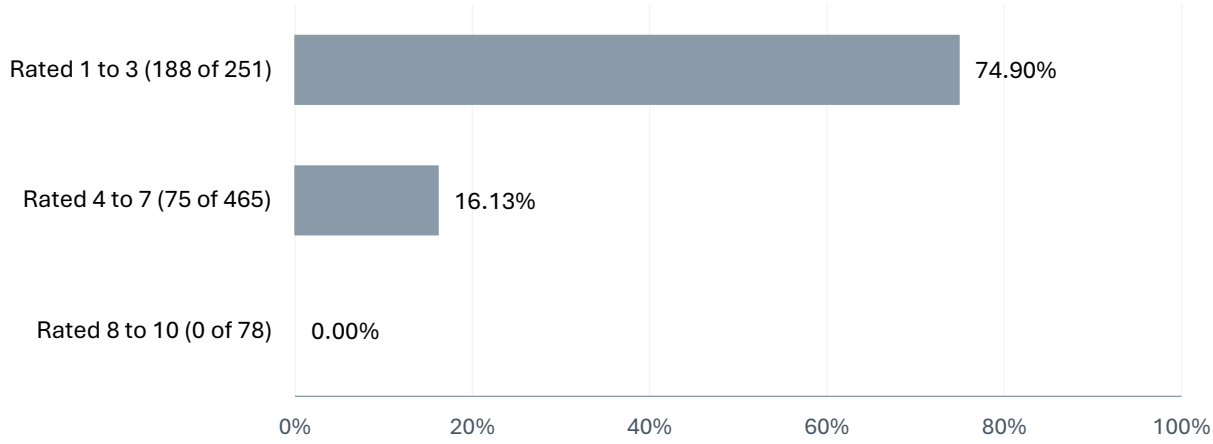
2.2 Monitoring for Unannounced Third-Party Change

Indicator C183 · Rate the engagement's monitoring for unannounced changes in the third-party APIs, data feeds, and models its agents depend on · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 7.12 · Full statistical analysis ›

Weak monitoring for unannounced provider change was associated with low scores. This control asks whether the engagement can detect changes in the interfaces, data feeds, and models on which agents depend. It extends oversight beyond a reliance on release notices. The adjusted model did not detect a clear separate cross-category association, while the written accounts supply context about how changes become visible.

Figure 3.66. Monitoring for Unannounced Third-Party Change: share with a score of 3 or below by indicator rating.

Weak monitoring for unannounced provider change was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 251 respondents who rated monitoring 1 to 3, 188 had a score of 3 or below. Among the 465 who rated it 4 to 7, 75 (16.13%) did, and none of the 78 who rated it 8 to 10 did (chi-square = 297.02, $p < 0.001$, Cramér's V = 0.61).

Monitoring was associated with the score computed without it at $r = 0.84$ (95% CI 0.79 to 0.88, $n = 150$ engagements) and with governance in the other four categories at $r = 0.76$. After adjustment for the mean of the other 53 indicators, the partial r was −0.12 ($p = 0.136$). Respondents who rated it distressed recorded a score of 2.40 ($n = 251$), against 4.72 ($n = 543$) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at $r = 0.69$ and across the active stages at 0.71 to 0.75, and after Holm adjustment it held in 84 of the 87 segment levels

tested. Within role it stood at 0.81 among individual contributors, 0.76 among engagement managers, and 0.75 among portfolio directors. Among individual contributors alone, 77.56% (159 of 205) of those who rated it 1 to 3 had a score of 3 or below.

Forty-six respondents on 39 engagements wrote about monitoring, and 30 of the answers were negative, 12 from canceled and 18 from active engagements. Of the 30, 22 described one of the 26 forms of a control bypassed or not working, 21 of them change discovered only through failure. Respondents described no alarm on the third-party side, so that employees reported a failure before any system did. Where monitoring had been built, it caught a change within the hour and left the work it disrupted to be done manually.

> *"Ask anyone here, the third party side has no alarm on it, we find out when the bot start failing, employees tell us before any system do"*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Corporate and back-office functions

> *"Our transcription provider changed its speaker labelling format in June without telling anyone, and our own monitoring caught it inside forty minutes, which is why this block scores well from me. What ratings do not show is what came after, two analysts labelled calls by hand for a day and a half while mapping was corrected, and if volume had been three times larger we simply could not have done it. Monitoring here is genuinely strong, contracts with these providers are much weaker, none of them owes us notice of anything."*
>
> **ENGAGEMENT MANAGER**
> CXM, Credit, underwriting and suitability

Part 2 (Third-Party Ecosystem, section 2.2) reports the association of monitoring with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 7.12) its association with manual alteration.

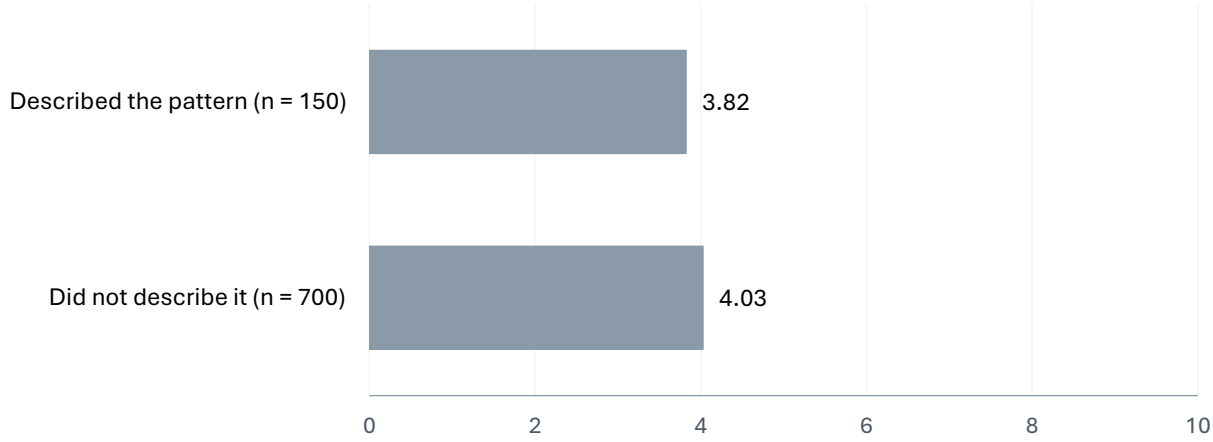
2.3 Changes Discovered Only Through Failure

Written pattern · All written answers to the open questions · 157 answers from 150 respondents on 93 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 7.7

Discovering provider change through failure was reported across a range of scores. Detection only after disruption leaves the engagement responding after a provider change has affected operations. Its written reach and the score comparisons describe different questions, so the number of engagements mentioning the pattern cannot establish its full incidence. The lack of a clear adjusted association does not make late discovery an adequate control arrangement.

Figure 3.67. Changes Discovered Only Through Failure: descriptive mean score comparison.

Discovering provider change through failure was reported across a range of scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 150 respondents who described it stood at 3.82 and the other 700 at 4.03 (cluster-robust p = 0.207). Among individual contributors alone, the 113 who described it stood at 3.73 against 3.89 among the other 554 (cluster-robust p = 0.369).

At the engagement level, the share of respondents describing the pattern was associated with the score at r = −0.17 (p = 0.042, n = 150 engagements). After adjustment for the mean of the other 53 indicators the association fell to partial r = −0.04 (p = 0.657). The adjusted comparison did not detect a clear association beyond the comparison mean.

The pattern reached every numeric score band. It was described by 20.14% (57 of 283) of respondents at 3 or below, 16.48% (87 of 528) above 3 to 7, and 6 of the 39 above 7. It was

described by 15.43% of respondents in Build-Out, 14.80% in Transition/Cut-Over, 13.59% in Production, and 22.36% on canceled engagements, and by 22.42% in the lowest indicator quintile against 17.50% in the highest. Individual contributors described it at 16.94% (113 of 667), engagement managers at 20.67% (31 of 150), and portfolio directors at 6 of 31, with no clearly detected difference by role after Holm adjustment. No segment family differed after Holm adjustment.

The 157 answers came from all written answers to the open questions, and 76 of them were written on canceled engagements. Part 5 counts 26 answers from 25 respondents on 22 engagements for the same pattern, among the written answers on the manual alteration methods alone, so the two counts rest on different bases. Respondents described runs that failed on a Monday after a weekend release and queues that went quiet. On the highest-scoring engagement of the study, a respondent reported that the providers' change notes arrived by email and went unread until something broke.

> *The rate limit is the thing people complain about but it is not really the providers fault, we send the whole book in one go at month end instead of spreading it, any provider will throttle that. Their change notes come by email every release also, we simply do not read them until something break. On the licence question i put 7 but that is because nobody has asked them yet, not because they refused"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Credit, underwriting and suitability

> *On my previous account the feed providers sent change notices weeks ahead. Here we normally find out because a tick handler has already stopped"*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Technology engineering and operations

Part 5 (Leading Dynamics of Manual Alteration, section 7.7) analyzes the pattern as a factor in manual alteration.

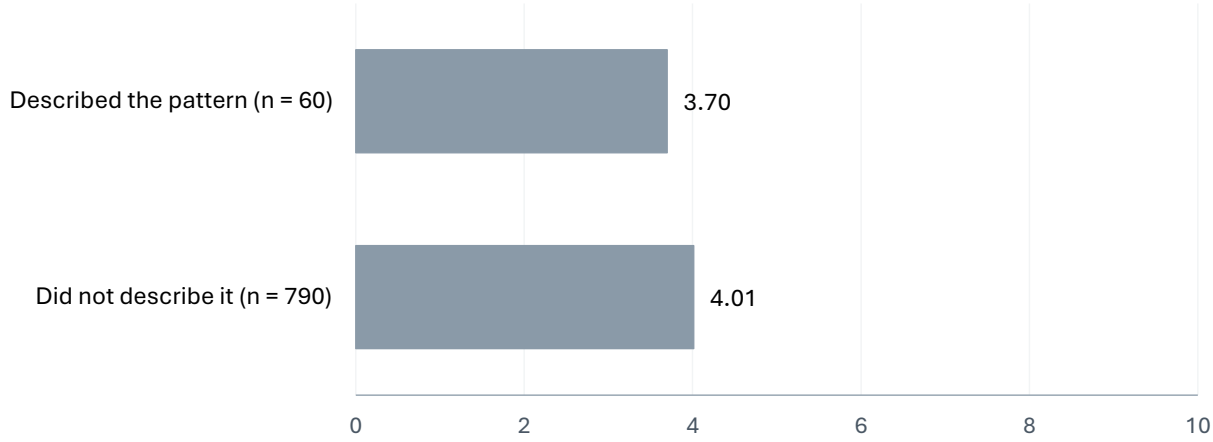
2.4 Schema and Format Changes Without Notice

Written pattern · All written answers to the open questions · 61 answers from 60 respondents on 51 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 7.6

Unannounced schema and format changes were described without a clearly detected difference in respondent mean scores. This identifies a specific external-change problem that a broad composite may not sharply distinguish. The governance question concerns preparation for changes to the structure agents expect, separate from general provider notice. The evidence remains a set of written descriptions, with no verified nonincident group among those who did not mention it.

Figure 3.68. Schema and Format Changes Without Notice: descriptive mean score comparison.

Unannounced schema and format changes were described without a clearly detected difference in respondent mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 60 respondents who described it stood at 3.70 and the other 790 at 4.01 (cluster-robust p = 0.187). Among individual contributors alone, the 51 who described it stood at 3.72 against 3.88 among the other 616 (cluster-robust p = 0.554).

At the engagement level, the share describing the pattern was associated with the score at r = −0.16 (p = 0.045, n = 150 engagements), and after adjustment for the mean of the other 53 indicators the association fell to partial r = −0.06 (p = 0.478).

The pattern was described by 8.13% (23 of 283) of respondents with a score of 3 or below, 6.63% (35 of 528) above 3 to 7, and 2 of the 39 above 7. It was described by 5.71% of

respondents in Build-Out, 5.20% in Transition/Cut-Over, 6.80% in Production, and 9.32% on canceled engagements, and by 8.48% in the lowest indicator quintile against 6.25% in the highest. Individual contributors described it at 7.65% (51 of 667), engagement managers at 4.00% (6 of 150), and portfolio directors at 3 of 31, with no clearly detected difference by role after Holm adjustment. No segment family differed after Holm adjustment.

The 61 answers came from all written answers to the open questions, and 31 of them were written on canceled engagements. Part 5 counts 27 answers from 26 respondents on 24 engagements, among the written answers on the manual alteration methods alone, so the two counts rest on different bases. Respondents described a renamed column, a changed date format, or an inserted field that failed every downstream mapping at once and surfaced only at reconciliation.

> *“Fund administrator change their file layout in June, no notice given. We found it when reconciliation failed... Nobody monitor for this, we just see it break.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Trade, settlement and reconciliation

> *“supplier risk feed changed the date format one week... nobody tell us anything. agents dropped every record and i only notice because count was too low. no email, no note, we find out ourselves”*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Corporate and back-office functions

Part 5 (Leading Dynamics of Manual Alteration, section 7.6) analyzes the pattern as a factor in manual alteration.

3 Contractual and Licensing Coverage of Agent Work

Contractual coverage of disruptive provider change was associated with the wider score. This addresses a formal boundary over external dependency, while the notice and monitoring indicators address operation. The adjusted association was not clearly detected. The findings should distinguish reported contract coverage from evidence that contractual terms delivered effective operational protection.

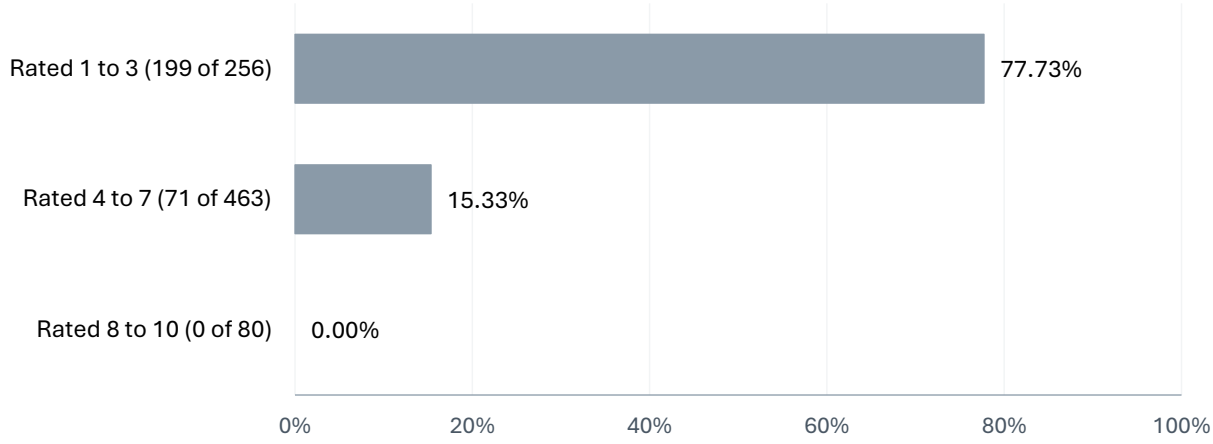
3.1 Contractual Coverage of Disruptive Provider Change

Indicator C184 · Rate how well the contracts with third-party providers deal with provider changes that can break the agents on this engagement · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 7.9 · Full statistical analysis ›

Weak contractual coverage of disruptive provider change was associated with low scores. Contracts define an external responsibility boundary, but their reported coverage does not by itself show that operational notice, monitoring, or response works. The relatively higher mean rating and the large low-score share among low indicator ratings concern different quantities. Both should inform a restrained assessment of contractual control.

Figure 3.69. Contractual Coverage of Disruptive Provider Change: share with a score of 3 or below by indicator rating.

Weak contractual coverage of disruptive provider change was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

The mean rating stood at 4.66, level with the reliability of guardrails. Of the 256 respondents who rated contractual coverage 1 to 3, 199 had a score of 3 or below, the second-highest share of the ten. Among the 463 who rated it 4 to 7, 71 (15.33%) did, and none of the 80 who rated it 8 to 10 did (chi-square = 332.27, $p < 0.001$, Cramér's V = 0.64).

Contractual coverage was associated with the score computed without it at $r = 0.84$ (95% CI 0.79 to 0.88, n = 150 engagements) and with governance in the other four categories at $r = 0.76$. After adjustment for the mean of the other 53 indicators, the partial r was −0.07 ($p = 0.406$). Respondents who rated it distressed recorded a score of 2.27 (n = 256), against 4.77 (n = 543) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at $r = 0.71$ and across the active stages at 0.71 to 0.76, and after Holm adjustment it held in all 86 segment levels tested. Within role it stood at 0.82 among individual contributors, 0.78 among engagement managers, and 0.70 among portfolio directors. Among individual contributors alone, 78.87% (168 of 213) of those who rated it 1 to 3 had a score of 3 or below.

Twenty-one respondents on 21 engagements wrote about contracts, and 14 of the answers were negative, 6 from canceled and 8 from active engagements. Four of the 14 described one of the 26 forms of a control bypassed or not working. Respondents described a delivery team that held no contract with the providers its agents depended on, and contract terms that did not oblige a provider to give notice of change.

> *“We do not hold the contract with the HR feed provider, that is the part a score cannot show. Their directory sync breaks, leavers stay active, and all we can do is tell the bank and wait.”*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Technology engineering and operations

> *“Dependencies are the source control platform, the artefact registry, the ticketing system and one model provider. Licensing was checked at contract stage and covers agent use, so no exposure there. Contract terms with the registry vendor do not oblige them to notify us of API changes. There is no fallback registry configured. If it goes down, releases stop until it returns; we have not tested how long that would take. Nothing to add on the rest.”*
>
> **PORTFOLIO DIRECTOR**
>
> ITO, Technology engineering and operations

Part 2 (Third-Party Ecosystem, section 4.1) reports the association of contractual coverage with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 7.9) its association with manual alteration.

“

The full list of our dependencies has never been written down. I know our own test tools and the model behind them, but under that there are libraries and one data service the client environment calls, and I could not tell you who owns those. Last month a licence expired on a plug-in and the whole nightly run failed, nobody knew until morning.”

INDIVIDUAL CONTRIBUTOR

ITO, Technology engineering and operations

4 Security Posture and Provider Concentration

Third-party security posture and provider concentration were associated with the wider score. They concern different exposures: the condition of participating services and dependence on a limited set of providers. The written base is smaller than the rating base. Security posture concerns the condition of participating services, while concentration concerns the resilience of dependence on a limited set of providers. Comment counts do not measure their relative importance.

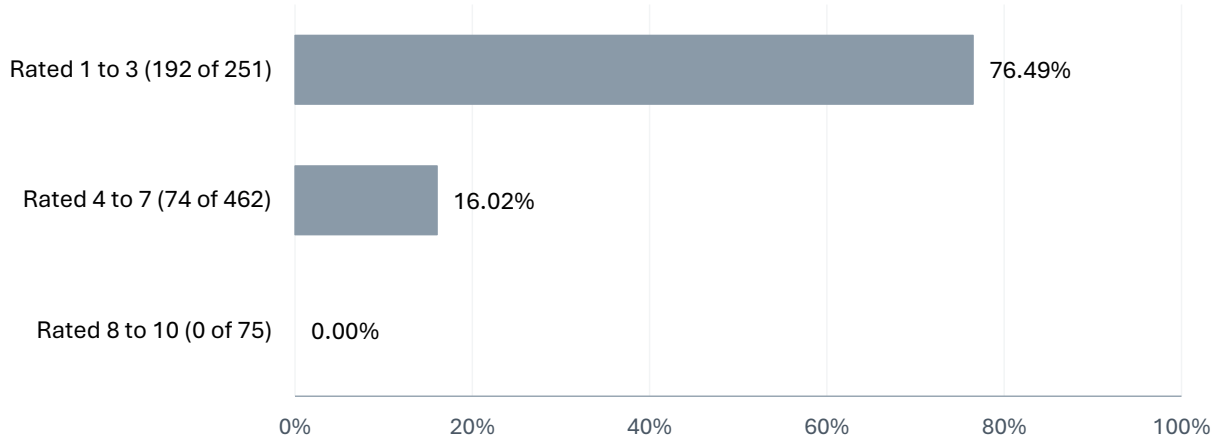
4.1 Security Posture of Third-Party Tools and Sub-Processors

Indicator C180 · Rate the security posture of the third-party tools, platforms, and sub-processors the engagement's agents connect to · Full statistical analysis ›

Poor reported security posture of third-party tools and sub-processors was associated with low scores. The indicator extends governance to parties whose systems or services participate in the agent workflow. Its written evidence is small and cannot establish the extent of security exposure. The stronger quantitative base remains respondents' ratings, which should not be represented as an independent security assessment.

Figure 3.70. Security Posture of Third-Party Tools and Sub-Processors: share with a score of 3 or below by indicator rating.

Poor reported security posture of third-party tools and sub-processors was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 251 respondents who rated security posture 1 to 3, 192 had a score of 3 or below. Among the 462 who rated it 4 to 7, 74 (16.02%) did, and none of the 75 who rated it 8 to 10 did (chi-square = 308.25, p < 0.001, Cramér's V = 0.63).

Security posture was associated with the score computed without it at r = 0.84 (95% CI 0.78 to 0.88, n = 150 engagements) and with governance in the other four categories at r = 0.76. After adjustment for the mean of the other 53 indicators, the partial r was −0.14 and was not significant (p = 0.083), the most negative of the ten. Respondents who rated it distressed recorded a score of 2.35 (n = 251), against 4.72 (n = 537) among the rest (cluster-robust p < 0.001).

The association was observed within the canceled engagements at r = 0.73, the strongest of the ten, and across the active stages at 0.69 to 0.76. After Holm adjustment, it was significant in all 86 segment levels tested. Within role it stood at 0.81 among individual contributors and 0.80 among engagement managers and portfolio directors. Among individual contributors alone, 76.67% (161 of 210) of those who rated it 1 to 3 had a score of 3 or below.

Nine respondents on 8 engagements wrote about security posture, and four of the answers were negative. Three of the four described one of the 26 forms of a control bypassed or not working, two of them problems raised with no response. The answers described a provider login shared by a whole team with no record of who used it, and a security review that stopped at the first layer of suppliers. An individual contributor added that the list of dependencies requested in the first month had never been written.

> *“Truthfully nobody sat down and wrote a list of what we depend on. I asked for it in month one and there is still no list. That is not really a third party fault though, they publish their notes, the problem is on our end that nobody reads them or tells us. The card scheme side actually gave notice about the sandbox refresh, it was in an email somewhere, only it went to a person who left. So I would not say the providers are bad. The security posture of the smaller tools we plugged in, that one I do worry about, we connected two or three things without anyone checking them properly and I raised it and got no answer. Also the licences, I do not think anyone thought about an agent making calls instead of a person, that came up when the quota looked strange and we had to discuss about it for days. But to be fair to everyone this account was only three months and half the access never came through, so there was never a moment where somebody could stop and map it all out. You cannot map what you have not been given. I think if we had another six months it would have been fine, or at least better... hard to say now.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Technology engineering and operations

Part 2 (Third-Party Ecosystem, section 6.1) reports the association of security posture with the four study measures.

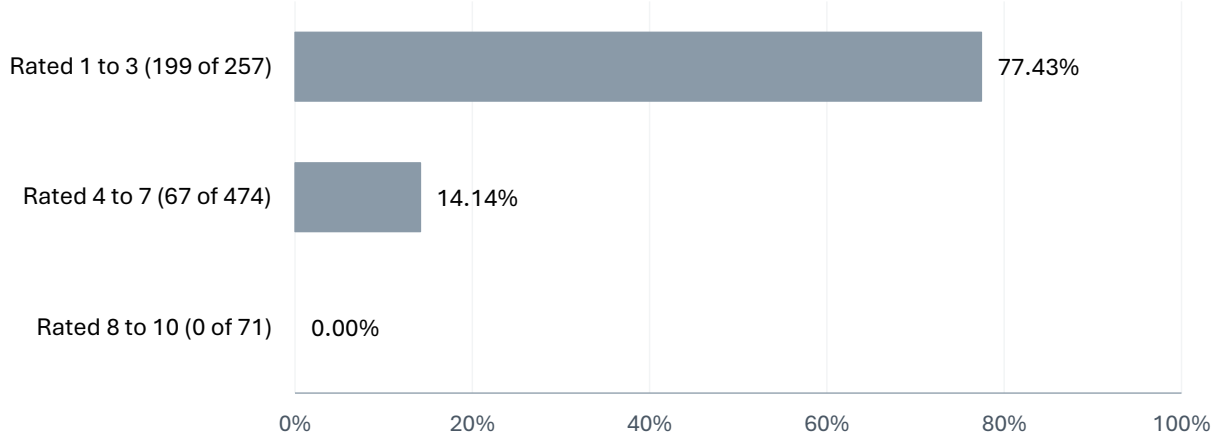
4.2 Management of Dependence on a Small Number of Model and Cloud Providers

Indicator C193 · Rate how well the engagement manages its dependence on a small number of model and cloud providers · Full statistical analysis ›

Weak management of dependence on a small number of model and cloud providers was associated with low scores. This concerns concentration of operational dependency, distinct from knowing which providers exist in the chain. The control question is whether that dependence is managed within the engagement's governance arrangements. The findings do not establish the probability or consequences of a provider disruption.

Figure 3.71. Management of Dependence on a Small Number of Model and Cloud Providers: share with a score of 3 or below by indicator rating.

Weak management of dependence on a small number of model and cloud providers was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 257 respondents who rated the management of concentration 1 to 3, 199 had a score of 3 or below, the third-highest share of the ten. Among the 474 who rated it 4 to 7, 67 (14.14%) did, the lowest such share of the ten, and none of the 71 who rated it 8 to 10 did (chi-square = 339.86, $p < 0.001$, Cramér's V = 0.65).

The indicator was associated with the score computed without it at $r = 0.84$ (95% CI 0.79 to 0.88, n = 150 engagements) and with governance in the other four categories at $r = 0.76$. After adjustment for the mean of the other 53 indicators, the partial r was −0.04 ($p = 0.646$), the

nearest to zero of the ten. Respondents who rated it distressed recorded a score of 2.32 (n = 257), against 4.78 (n = 545) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at $r = 0.72$ and ranged across the active-stage groups from 0.73 in Build-Out to 0.77 in Production. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.81 among individual contributors and engagement managers and 0.80 among portfolio directors. Among individual contributors alone, 77.83% (165 of 212) of those who rated it 1 to 3 had a score of 3 or below.

Twenty-six respondents on 24 engagements wrote about provider concentration, and 19 of the answers were negative, 7 from canceled and 12 from active engagements. Four of the 19 described one of the 26 forms of a control bypassed or not working. Respondents described a single model provider with no second path, and a retirement notice that the engagement detected in time and could not act on.

> *“Only real worry is we are on one cloud provider for everything, if that region goes we have nothing, rest of the feeds have been steady enough”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

Part 2 (Third-Party Ecosystem, section 6.2) reports the association of the management of concentration with the four study measures.

5 Fallback Capacity and Retention of Operational Knowledge

Converting informal human fixes into explicit checks was associated with the wider score. This connects operational knowledge with the visibility and repeatability of control after agent work is introduced. The separately coded habit pattern was not clearly associated with the rating. That result qualifies the measurement relationship without demonstrating that personal knowledge or fallback arrangements were adequate.

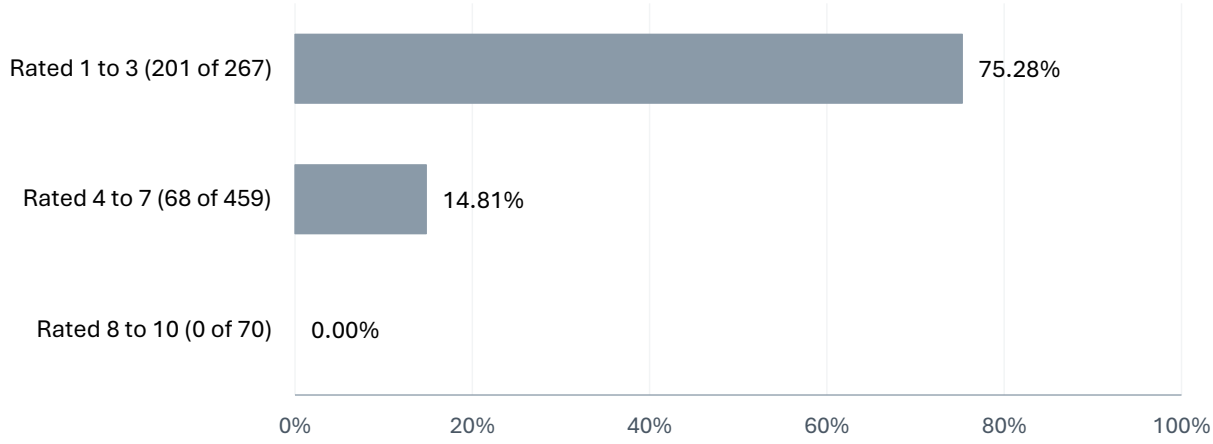
5.1 Replacement of Informal Human Fixes With Explicit Checks

Indicator C192 · Rate how deliberately the quiet human fixes for third-party quirks were replaced with explicit checks when agents took over the work · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 7.16 · Full statistical analysis ›

Failure to convert informal human fixes into explicit checks was associated with low scores. The indicator concerns whether knowledge previously carried by people becomes a visible and repeatable part of agent-era control. The separately coded habit pattern was not clearly associated with the rating. That comparison raises a measurement and interpretation question but does not formally demonstrate that the indicator cannot detect the practice.

Figure 3.72. Replacement of Informal Human Fixes With Explicit Checks: share with a score of 3 or below by indicator rating.

Failure to convert informal human fixes into explicit checks was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 267 respondents who rated the replacement of informal fixes 1 to 3, 201 had a score of 3 or below. Among the 459 who rated it 4 to 7, 68 (14.81%) did, and none of the 70 who rated it 8 to 10 did (chi-square = 315.03, $p < 0.001$, Cramér's V = 0.63).

The indicator was associated with the score computed without it at $r = 0.85$ (95% CI 0.80 to 0.89, n = 150 engagements) and with governance in the other four categories at $r = 0.77$. After adjustment for the mean of the other 53 indicators, the partial r was −0.11 ($p = 0.183$). Respondents who rated it distressed recorded a score of 2.37 (n = 267), against 4.76 (n = 529) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at $r = 0.72$ and ranged across the active-stage groups from 0.70 in Build-Out to 0.81 in Production. After Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.82 among individual contributors, 0.80 among engagement managers, and 0.76 among portfolio directors. Among individual contributors alone, 78.34% (170 of 217) of those who rated it 1 to 3 had a score of 3 or below.

Habit in place of a control did not follow the rating. It was described by 10.11% (27 of 267) of respondents who rated the replacement of informal fixes 1 to 3, 10.24% (47 of 459) of those who rated it 4 to 7, and 14.29% (10 of 70) of those who rated it 8 to 10 (chi-square $p = 0.567$). Informal fixes persisted as habit on engagements where respondents judged them to have been replaced.

Nine respondents on 9 engagements wrote about informal fixes, and 7 of the answers were negative. Four of the 7 described one of the 26 forms of a control bypassed or not working, two of them habit in place of a control. Respondents described manual checks that had never been written down when the agents took over, and an engagement manager observed that payroll corrections had lived in one supervisor's memory for eleven years.

> *“directory sync and the ticketing api have been steady, no complaint from me. one thing not asked, the small manual checks we used to do on stale accounts is not written anywhere, so when agent take over that step nobody know it existed”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

> *“The question about where people used to quietly fix things assumes somebody wrote it down. Nobody did. Payroll deduction corrections lived in one supervisors head for eleven years. You cant score that from one to ten. You can only find out after she leaves.”*
>
> **ENGAGEMENT MANAGER**
>
> BPO, Corporate and back-office functions

Part 2 (Third-Party Ecosystem, section 7.4) reports the association of the indicator with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 7.16) its association with manual alteration.

“

The vendor rolled the non-compliant investor letter into a monthly 'draft accuracy' composite score that averaged it against hundreds of routine internal messages, so the single high-risk failure barely moved the number the client saw.”

INDIVIDUAL CONTRIBUTOR

CXM, Client servicing and communications

6 The Dependency Chain and Gated External Rails

Dependency mapping and management of gated external rails were associated with the wider score. Mapping concerns knowledge of the full chain, while gated rails concern constraints imposed by systems such as payment networks and regulatory submissions. The distinction matters for governance because knowing a dependency does not establish authority to change it. Conditional percentage rankings do not measure the comparative importance of those responsibilities.

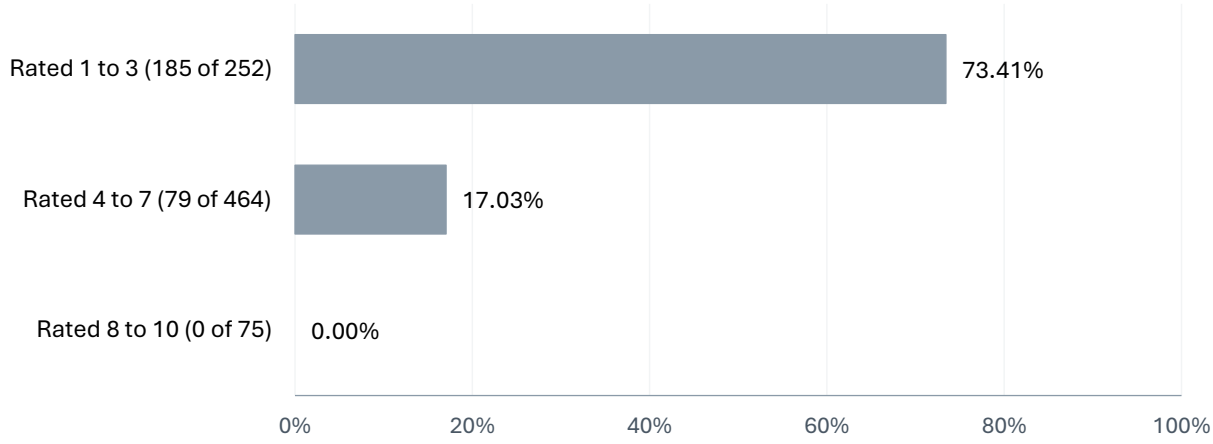
6.1 Mapping of Third- and Fourth-Party Dependencies

Indicator C196 · Rate how completely the engagement has mapped every third and fourth party its agents depend on, down the full chain · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 7.13 · Full statistical analysis ›

Incomplete mapping of third- and fourth-party dependencies was associated with low scores. Oversight requires visibility through the full chain of services on which agents depend. Mapping is a condition for understanding the chain, distinct from the ability to control every party within it. The adjusted comparison did not detect a clear separate association, and the conditional percentage ranking does not establish relative control importance.

Figure 3.73. Mapping of Third- and Fourth-Party Dependencies: share with a score of 3 or below by indicator rating.

Incomplete mapping of third- and fourth-party dependencies was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 252 respondents who rated dependency mapping 1 to 3, 185 had a score of 3 or below. Among the 464 who rated it 4 to 7, 79 (17.03%) did, the highest share at those ratings of the 47 governance indicators, and none of the 75 who rated it 8 to 10 did (chi-square = 275.01, $p < 0.001$, Cramér's V = 0.59).

Dependency mapping was associated with the score computed without it at $r = 0.84$ (95% CI 0.79 to 0.88, n = 150 engagements) and with governance in the other four categories at $r = 0.76$. After adjustment for the mean of the other 53 indicators, the partial r was −0.13 ($p = 0.105$). Respondents who rated it distressed recorded a score of 2.43 (n = 252), against 4.72 (n = 539) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at $r = 0.66$ and ranged across the active-stage groups from 0.71 in Build-Out to 0.78 in Production. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.81 among individual contributors and 0.78 among engagement managers and portfolio directors. Among individual contributors alone, 76.59% (157 of 205) of those who rated it 1 to 3 had a score of 3 or below.

Forty-four respondents on 37 engagements wrote about dependency mapping, and 35 of the answers were negative, 21 from canceled and 14 from active engagements. Of the 35, 13 described one of the 26 forms of a control bypassed or not working, most often change discovered only through failure (7). Respondents described engagements that knew their direct providers and no list beyond the first tier, including engagements where contracts had been tightened and the map had not been drawn.

> *“Two months in, contracts with the market data providers have been reviewed and they now carry notice terms for interface changes, we did not have that at kickoff. What has not been done is the dependency map. We know our direct providers, we dont know who sits behind them, and nobody has looked at whether our licences cover machine users at all.”*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Technology engineering and operations

Part 2 (Third-Party Ecosystem, section 8.1) reports the association of dependency mapping with the four study measures, and Part 5 (Leading Dynamics of Manual Alteration, section 7.13) its association with manual alteration.

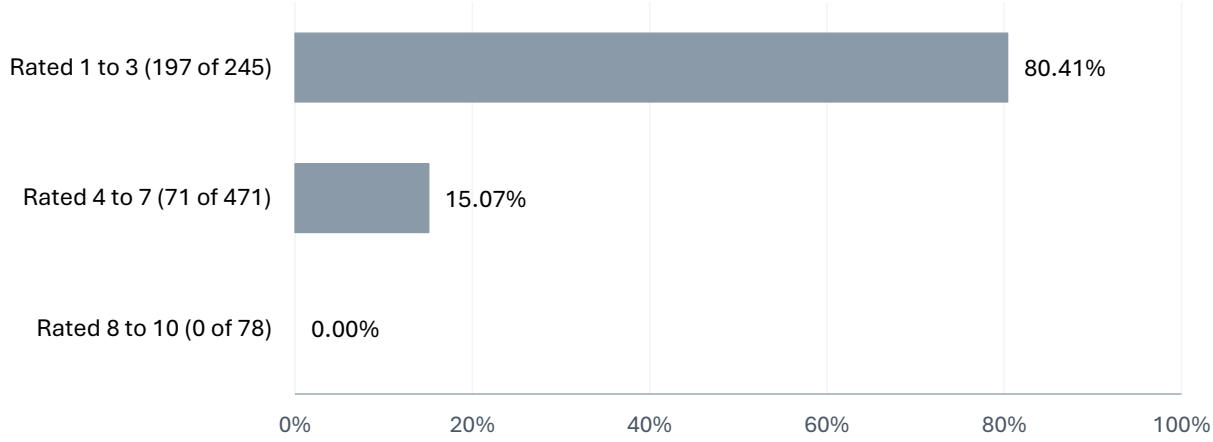
6.2 Management of Dependence on Gated External Rails

Indicator C197 · Rate how well the engagement manages agent dependence on gated external rails such as payment networks and regulatory submission systems · Full statistical analysis ›

Weak management of dependence on gated external rails was associated with low scores. Payment networks and regulatory submission systems can impose boundaries that the two firms cannot change through internal agent design. The governance question is how the engagement recognizes and manages those constraints. A large low-score share among low ratings is a descriptive association, not a validated measure of diagnostic reliability.

Figure 3.74. Management of Dependence on Gated External Rails: share with a score of 3 or below by indicator rating.

Weak management of dependence on gated external rails was associated with low scores.



Note: The score excludes this indicator. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the 245 respondents who rated the management of gated rails 1 to 3, 197 had a score of 3 or below. Among the 471 who rated it 4 to 7, 71 (15.07%) did, and none of the 78 who rated it 8 to 10 did (chi-square = 351.73, $p < 0.001$, Cramér's V = 0.67).

The indicator was associated with the score computed without it at r = 0.84 (95% CI 0.78 to 0.88, n = 150 engagements) and with governance in the other four categories at r = 0.76. After adjustment for the mean of the other 53 indicators, the partial r was −0.08 ($p = 0.333$). Respondents who rated it distressed recorded a score of 2.27 (n = 245), against 4.76 (n = 549) among the rest (cluster-robust $p < 0.001$).

The association was observed within the canceled engagements at r = 0.69 and across the active stages at 0.79 in Build-Out, 0.72 in Transition/Cut-Over, and 0.78 in Production. After

Holm adjustment, it was significant in 86 of the 87 segment levels tested. Within role it stood at 0.82 among individual contributors, 0.78 among engagement managers, and 0.71 among portfolio directors. Among individual contributors alone, 81.68% (165 of 202) of those who rated it 1 to 3 had a score of 3 or below, the highest such share of the 47.

Fourteen respondents on 12 engagements wrote 15 answers about gated rails, and 13 were negative, 8 from canceled and 5 from active engagements. Two of the 13 described problems raised with no response. Respondents described certification and filing portals built for a person holding a token, a matching utility that opened twice a day, and engagements where notice from the providers had improved while access to the rails had not.

> *My low marks are really about matching utility... it opens for us twice a day only, and if we miss window there is nothing else, whole batch waits till tomorrow”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Trade, settlement and reconciliation

> *Weakest point here is that some of what our agents must touch is still a screen designed for a person, particularly on the network side, and that is the one place I marked down. But I would not lay that at the provider's door as a failing. Card networks carry twenty years of member banks built around those screens, and an interface change there is not a release, it is a migration for several thousand institutions. They have given us honest notice of every change that mattered so far, in one case three months of it, which is more than we manage internally between our own teams. The fragility people speak about in this area is mostly ours, not theirs. We built agents on top of somebody else's roadmap and then treat the roadmap as a risk. I would rather we planned for their calendar than complained about it.”*
>
> **PORTFOLIO DIRECTOR**
>
> BPO, Payments, cards and disputes

Part 2 (Third-Party Ecosystem, section 8.2) reports the association of the management of gated rails with the four study measures.

Understanding Governing Obligations and the Control Environment

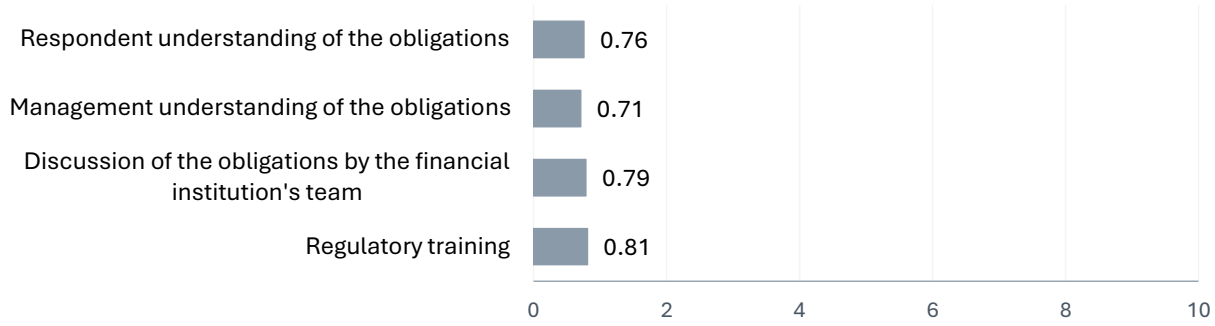
Reported understanding and discussion of governing obligations were associated with higher governance, oversight, and control scores. Merely naming a law did not clearly distinguish the score, and training's adjusted association was sensitive to the comparison measure. These findings separate recognition of an obligation, perceived understanding, operational discussion, and preparation to apply it. They concern reported knowledge and practice, without an independent test of legal compliance.

The respondent's own understanding of the obligations retained an association with the score after adjustment for the comparison mean outside it at partial r = 0.27, management's understanding at 0.25, and the financial institution's discussion of the obligations at 0.20. Regulatory training had a negative adjusted coefficient of −0.35 that changed sign when the vendor capability items were omitted from the comparison mean. Respondents who named at least one governing law or regulation recorded a score of 4.06 (n = 548), against 3.90 among the 247 who answered that they did not know. Written references to a regulation, a regulator, or an audit obligation were rare, at 16 answers from 14 respondents on 12 engagements. Table 3.13 sets out the four rated items by rating band.

Table 3.13. The score by rating of the four items on the governing obligations

Regulatory item	Mean rating (n)	Distressed ratings	The score, rating 1 to 3 (n)
Respondent understanding of the obligations	4.81 (800)	228 of 800 (28.50%)	2.41 (228)
Management understanding of the obligations	4.93 (516)	149 of 516 (28.88%)	2.42 (149)
Discussion of the obligations by the financial institution's team	4.78 (801)	231 of 801 (28.84%)	2.39 (231)
Regulatory training	3.58 (798)	385 of 798 (48.25%)	2.74 (385)

r with the score



Regulatory item	Rating 4 to 7 (n)	Rating 8 to 10 (n)	r with the score
Respondent understanding of the obligations	4.31 (484)	6.21 (88)	0.76
Management understanding of the obligations	4.40 (297)	6.16 (70)	0.71
Discussion of the obligations by the financial institution's team	4.35 (481)	6.23 (89)	0.79
Regulatory training	5.03 (402)	7.12 (11)	0.81

Note: The fifth item, the laws and regulations respondents named as governing the engagement, is a written answer and carries no rating. Respondents who named at least one law or regulation recorded a score of 4.06 (n = 548), against 3.90 among the 247 who answered I do not know. The 265 respondents who answered I do not know on management understanding are excluded from its bands. Regulatory items are not among the 47 governance indicators. The score is computed in full. These items measure reported understanding, discussion, and training, not independently established compliance. Ordinary respondent-level band-test p values do not account for engagement clustering. Adjusted-correlation p values are nominal.

> *We were so busy fixing pre-prod issues that we never thought about how to do the actual work."*
>
> **ENGAGEMENT MANAGER**
>
> KPO, Corporate and back-office functions

“

What they gave us for training was two weeks of generic e-learning and nothing on sanctions or aml typologies at all. i came in not knowing what a structuring pattern looks like and i learned it from the analyst sitting next to me. kindly note this is the whole reason my numbers here are low. everything else i can live with”

INDIVIDUAL CONTRIBUTOR

BPO, Financial crime and client onboarding

1 Knowledge of the Governing Obligations

The obligations items distinguish recognition, understanding, and reported exposure. Self-assessed understanding was associated with the score, whereas naming a law did not clearly distinguish it. The few written exposure accounts came from lower-scoring respondents. Together these findings connect knowledge with the reported control environment while limiting any inference that a named obligation or favorable understanding rating establishes compliance.

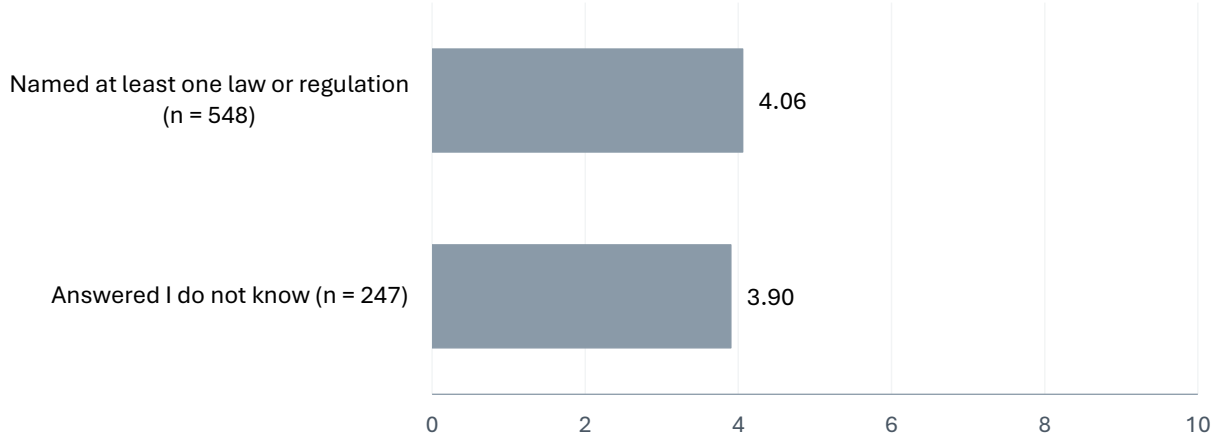
1.1 Governing Laws and Regulations Named by Respondents

Item C056 · The laws and regulations that govern the engagement, named in writing · 548 of 795 respondents named at least one

Naming a governing law or regulation did not clearly distinguish the reported control environment. The ability to name an obligation is different from understanding its operational implications or applying an effective control. The role pattern is also material, because all respondents answering that they did not know were individual contributors. The result does not establish equivalent legal knowledge or compliance among the groups.

Figure 3.75. The score of respondents who named a governing law or regulation and of those who did not know

Naming a governing law or regulation did not clearly distinguish the reported control environment.



Of the 795 respondents who answered, 548 (68.93%) named at least one law or regulation, on all 150 engagements. They recorded a score of 4.06, against 3.90 among the 247 who answered that they did not know (cluster-robust p = 0.198).

At the engagement level, the share naming a law was associated with the score at r = 0.07 (p = 0.403, n = 150 engagements), and after adjustment for the mean of the other 53 indicators the partial r was 0.00 (p = 0.983).

The difference that existed lay in role. All 150 engagement managers, all 29 portfolio directors who answered, and the one customer management executive who answered named a law, while 368 of the 615 individual contributors did, so every answer of I do not know came from an individual contributor. Among individual contributors alone, those who named a law recorded a score of 3.87, against 3.90 among those who did not (cluster-robust p = 0.836). The share naming a law moved within a narrow range across the indicator quintiles, from 97 of 153 respondents in the lowest to 109 of 158 in the highest, and across engagement status, from 117 of 164 respondents in Build-Out to 197 of 297 on canceled engagements.

Knowledge of the governing law sat beside failed controls in the written answers. An individual contributor on a regulatory return described entity codes that arrived inconsistent from two systems and were normalized by the delivery team without anyone asking.

> *Happens most weeks, i do not always discuss about it with lead, just fix and revert back if they ask. And to be fair nobody has ever asked. The regulatory return data comes across with the entity codes inconsistent between the two systems they extract from, and if you load it straight the aggregation is wrong at the entity level. Which for a regulatory return is not a small thing. So we normalise it first. We have been normalising it first since the second month of the engagement. It is in nobody's process document. I have assumed my lead has passed it up because he is diligent about that sort of thing, though i do not actually know. What i do know is the returns that go out are right, and they are right because somebody here noticed the codes did not line up. Also if that person moves on, the next person may not notice, and the return would still go out, just wrong. That is the thing that i think about occasionally. Not the work itself, the work is fine, it is more that the"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Regulatory, audit, tax and legal

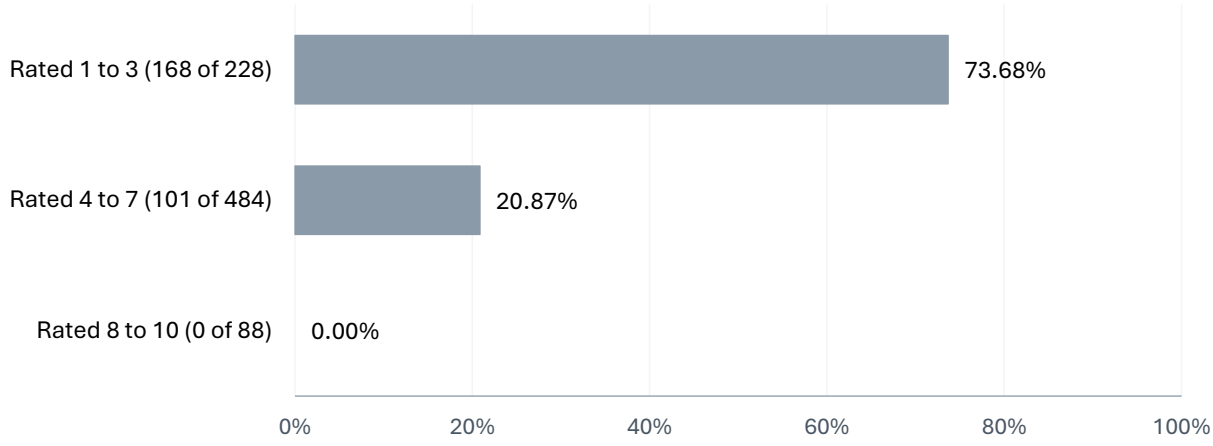
1.2 Respondent Understanding of the Governing Obligations

Survey item C057 · I fully understand how these laws, regulations, and similar affect the outsourcing engagement · Rated item analyzed only in Part 3

Greater self-reported understanding of governing obligations was associated with higher scores, including under the main adjustment. This connects awareness of the obligations with the reported control environment rather than treating a named law as evidence of governance. The measure remains a self-assessment. It does not independently test knowledge or demonstrate that the engagement complied with the obligations respondents had in mind.

Figure 3.76. Respondent Understanding of the Governing Obligations: share with a score of 3 or below by indicator rating.

Greater self-reported understanding of governing obligations was associated with higher scores, including under the main adjustment.



Note: The score uses all 47 governance indicators. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Understanding was associated with the score at r = 0.76 (95% CI 0.68 to 0.82, n = 150 engagements), and after that adjustment the association remained at partial r = 0.27 ($p < 0.001$), the strongest positive adjusted association of the four rated items.

Respondents who rated their understanding distressed recorded a score of 2.41 (n = 228), against 4.60 (n = 572) among the rest (cluster-robust $p < 0.001$). Of the 228, 168 (73.68%) had a score of 3 or below. Among the 484 who rated their understanding 4 to 7, 101 (20.87%) did, and none of the 88 who rated it 8 to 10 did (chi-square = 243.81, $p < 0.001$, Cramér's V = 0.55).

The association was observed within the canceled engagements at r = 0.61 and across the active stages at 0.64 to 0.71, and after Holm adjustment it held in 84 of the 86 segment levels tested. Within role it stood at 0.75 among individual contributors, 0.73 among engagement

managers, and 0.66 among portfolio directors. Among individual contributors alone, 73.44% (141 of 192) of those who rated their understanding 1 to 3 had a score of 3 or below.

Eight respondents on 6 engagements wrote about their understanding of the obligations, and only one of the answers was negative. Six of the seven favorable answers came from canceled engagements and named regulatory understanding among what the respondent took from an engagement that failed. One individual contributor on a canceled engagement listed the regulatory accountability question, learned in depth, among what the engagement had given.

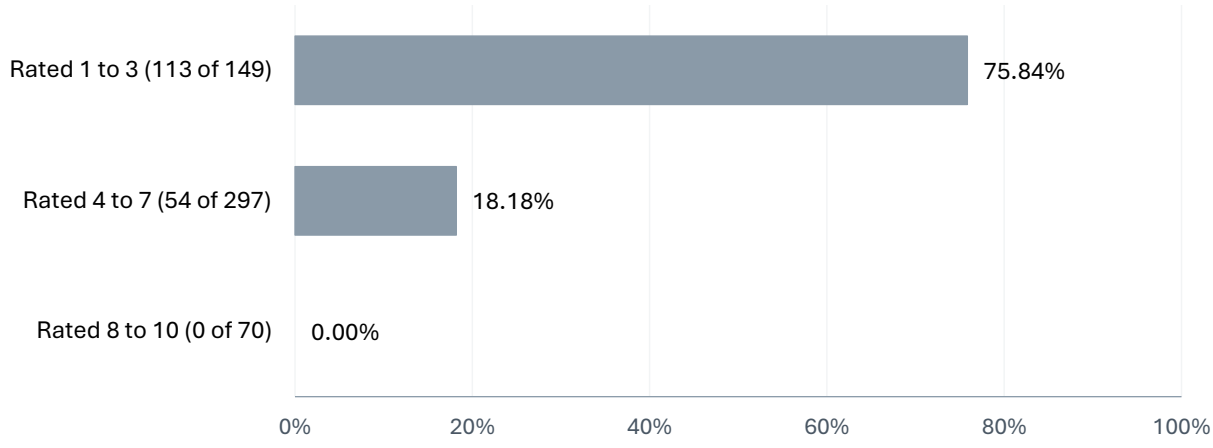
1.3 Management Understanding of the Governing Obligations

Survey item C058 · My management team fully understands these laws, regulations, and similar, and how they affect the engagement · Rated item analyzed only in Part 3

Respondents' assessments of management understanding were associated with the score, while many respondents could not provide a numeric assessment. This distinguishes management's reported knowledge from the workforce's visibility into that knowledge. Unknown answers identify uncertainty about management understanding and are excluded from the numeric rating comparison. They do not establish that management had low knowledge.

Figure 3.77. Management Understanding of the Governing Obligations: share with a score of 3 or below by indicator rating.

Respondents' assessments of management understanding were associated with the score, while many respondents could not provide a numeric assessment.



Note: The score uses all 47 governance indicators. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Of the respondents who answered, 265 answered that they did not know, and they recorded a score of 3.83. Among the respondents who gave a rating, management's understanding was associated with the score at r = 0.71 (95% CI 0.62 to 0.78, n = 150 engagements), and after adjustment for the mean of the other 53 indicators the association remained at partial r = 0.25 (p = 0.002).

Respondents who rated management's understanding distressed recorded a score of 2.42 (n = 149), against 4.73 (n = 367) among the rest (cluster-robust p < 0.001). Of the 149, 113 (75.84%) had a score of 3 or below, the highest such share of the four items. Among the 297 who rated it 4 to 7, 54 (18.18%) did, and none of the 70 who rated it 8 to 10 did (chi-square = 189.44, p < 0.001, Cramér's V = 0.61).

The association was observed within the canceled engagements at r = 0.63 and across the active stages at 0.66 to 0.69, and after Holm adjustment it held in 81 of the 84 segment levels tested. Within role it stood at 0.74 among individual contributors, 0.75 among engagement managers, and 0.71 among portfolio directors. Among individual contributors alone, 79.09% (87 of 110) of those who rated it 1 to 3 had a score of 3 or below.

No written answer addressed management's understanding of the obligations, and 265 respondents could not rate it, so the study holds no testimony on how management came to understand the obligations or failed to.

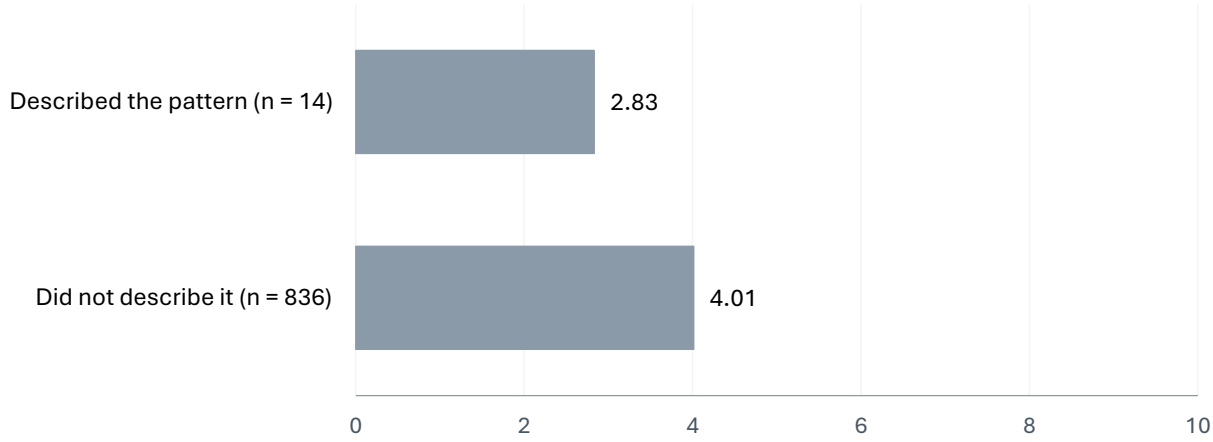
1.4 Regulatory or Compliance Exposure Reported by Respondents

Written pattern · All written answers to the open questions · 16 answers from 14 respondents on 12 engagements · Counted in Part 5, Manual Alteration by the Delivery Team, "Regulatory and Legal Context Reported by Respondents"

The small group describing regulatory or compliance exposure had lower mean scores. These accounts connect perceived exposure with the control environment, but they do not establish a legal breach or an exposure rate for the study population. The engagement-level association was not clearly detected. The appropriate interpretation preserves the concern described and the limits of the small written base.

Figure 3.78. Regulatory or Compliance Exposure Reported by Respondents: descriptive mean score comparison.

The small group describing regulatory or compliance exposure had lower mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

Mean scores were 2.83 for the 14 respondents describing exposure and 4.01 for the other 836 (cluster-robust p = 0.005). Among individual contributors alone, the 10 who described exposure stood at 2.65, against 3.88 among the other 657 (cluster-robust p = 0.008).

At the engagement level, the share describing exposure was associated with the score at r = −0.14 (p = 0.099, n = 150 engagements), and after adjustment for the mean of the other 53 indicators the partial r was −0.08 (p = 0.309). The 12 engagements were too few for the exposure to separate engagements by their governance.

Exposure was described by 8 of the 283 respondents with a score of 3 or below, 6 of the 528 above 3 to 7, and none of the 39 above 7. It appeared on canceled engagements and in every

stage except Transition/Cut-Over, and it reached 10 of the 667 individual contributors and 4 of the 150 engagement managers, with no clearly detected difference by role after Holm adjustment. No segment family differed after Holm adjustment.

The 16 answers came from all written answers to the open questions, and 9 of them were written on canceled engagements. Part 5 counts 20 answers on 14 engagements that referred to a regulation, a regulator, or an audit obligation, among the written answers on the manual alteration methods alone, so the two counts rest on different bases. Respondents described a regulatory deadline that would not wait for a rerun, a tool that misquoted regulatory guidance to a borrower in hardship, and compliance reports that cleared a non-compliant letter.

> *“the problem is same case, same data, different result next run, so i correct it manually before it goes to QA, this happen more than i like to admit on dispute side, you rerun same chargeback case and get different recommendation back, no warning, no explanation, and dispute case cannot have mistake in it because Reg E deadline does not care that model changed its mind, so i sit there and go through informations again myself before submitting, i raise it with my lead every time truthfully, we dont log it separate though, its just treated as part of normal QA now, feels like nobody upstream really want to hear about it anymore, i said it enough time already, card network side is no better, evidence file come in miscoded half the time and we wait weeks for any answer while deadline get closer and closer, so between that and model flipping on same case i end up doing double work most day, im tired of explaining same problem over and over to be honest, feedbacks from client escalation channel is slow too so even when we do flag something upstream nothing move fast, i just check output twice now before i trust it and if the two run disagree i take whichever one match what i already know about the case, its not ideal but dispute deadline dont move for anyone so”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Payments, cards and disputes

Part 5 (Manual Alteration by the Delivery Team, “Regulatory and Legal Context Reported by Respondents”) reports the regulatory references in the answers on manual alteration.

“

In fairness to the providers, these are proper machine interfaces, not screens we scrape, and the claims feeds arrive clean enough that the agents can work them without us in the middle. Where we are genuinely thin is our own fallback. If a model provider went dark on a Tuesday morning we would be running claim status by hand, and half the desk has never worked that way.”

ENGAGEMENT MANAGER

CXM, Client servicing and communications

2 Discussion and Training on the Governing Obligations

Discussion and training concern how obligations enter day-to-day governance. Reported discussion by the financial institution was positively associated with the score under the main adjustment. Training effectiveness showed a more model-dependent result. Their distinct patterns support separate treatment of communication and preparation, rather than a single conclusion about legal awareness.

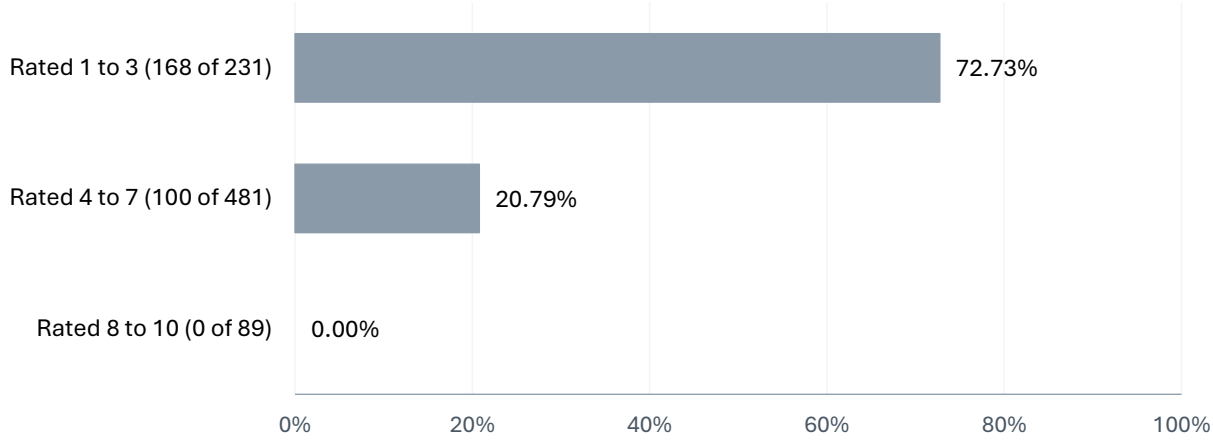
2.1 Discussion of the Obligations by the Financial Institution's Team

Survey item C059 · The customer's team actively discusses these laws, regulations, and similar with our team · Rated item analyzed only in Part 3

More active discussion of governing obligations by the financial institution's team was associated with higher scores. The indicator concerns whether obligations enter the joint operational conversation, beyond whether respondents can name or understand them individually. The positive adjusted estimate supports a conditional association with the control environment. Discussion itself remains distinct from evidence that the required controls were implemented effectively.

Figure 3.79. Discussion of the Obligations by the Financial Institution's Team: share with a score of 3 or below by indicator rating.

More active discussion of governing obligations by the financial institution's team was associated with higher scores.



Note: The score uses all 47 governance indicators. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Discussion was associated with the score at r = 0.79 (95% CI 0.72 to 0.84, n = 150 engagements), and after that adjustment the association remained at partial r = 0.20 (p = 0.017).

Respondents who rated the financial institution's discussion distressed recorded a score of 2.39 (n = 231), against 4.64 (n = 570) among the rest (cluster-robust p < 0.001). Of the 231, 168 (72.73%) had a score of 3 or below. Among the 481 who rated it 4 to 7, 100 (20.79%) did, and none of the 89 who rated it 8 to 10 did (chi-square = 239.42, p < 0.001, Cramér's V = 0.55).

The association held after Holm adjustment in all 85 segment levels tested, the only one of the four items to do so. It held within the canceled engagements at r = 0.65 and across the active stages at 0.67 to 0.71. Within role it stood at 0.78 among individual contributors and 0.74 among engagement managers and fell to 0.53 among portfolio directors. Among individual contributors alone, 75.38% (147 of 195) of those who rated it 1 to 3 had a score of 3 or below.

No written answer addressed the financial institution's discussion of the obligations directly. An engagement manager on a canceled regulatory engagement reported that the financial institution's compliance staff answered every call, while the delivery team could not learn which of three regional versions of the control the work was to follow.

> *“i felt for their compliance people to be candid. They answered every time we called. The problem was that surveillance runs three different ways in three regions over there and nobody could tell us which one we were supposed to build to.”*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Regulatory, audit, tax and legal

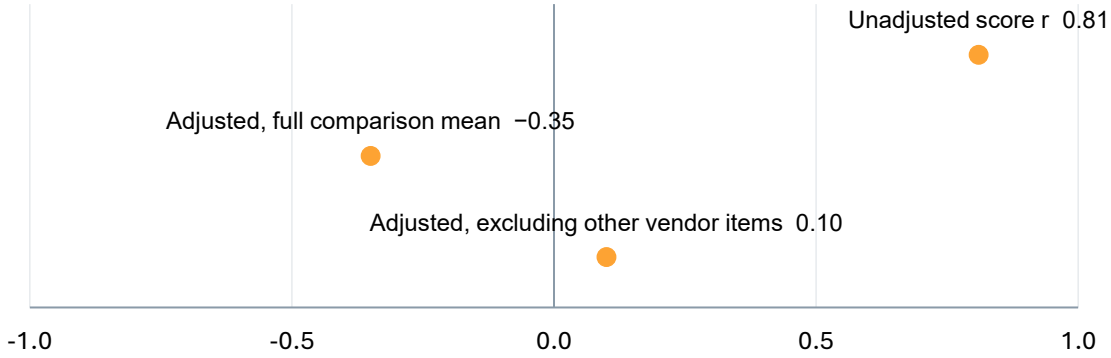
2.2 Regulatory Training

Survey item C060 · The training available to me effectively teaches me about these laws, regulations, and similar · Rated item analyzed only in Part 3

Reported effectiveness of regulatory training was associated with the wider score, but its adjusted coefficient was sensitive to the comparison measure. Training concerns whether the workforce is equipped to apply the obligations to its work. Its negative coefficient in the main adjusted model should not be read as training worsening governance. The category-removal result limits that interpretation and leaves effectiveness as a reported assessment.

Figure 3.80. Regulatory Training: observed association and adjustment sensitivity.

Reported effectiveness of regulatory training was associated with the wider score, but its adjusted coefficient was sensitive to the comparison measure.



Comparison	Estimate and analysis base
Unadjusted score r	r = 0.81 150 engagements
Adjusted, full comparison mean	r = −0.35 150 engagements
Adjusted, excluding other vendor items	r = 0.10 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

Correlations use 150 engagements. The alternative mean removes the stated category items. Estimates are descriptive model comparisons, with nominal inference in the text.

Training was distressed for 48.25% of the respondents who rated it, against 28.50% to 28.88% for the other three rated items.

Training was associated with the score at r = 0.81 (95% CI 0.74 to 0.86, n = 150 engagements), the highest of the four items. After adjustment for the mean of the other 53 indicators, the

partial r was −0.35 ($p < 0.001$). The result stayed inverse when the financial institution readiness block was omitted from the comparison mean (−0.25, $p = 0.002$), when the AI technology block left it (−0.30, $p < 0.001$), and when the third-party block left it (−0.24, $p = 0.004$). It reversed to 0.10 ($p = 0.232$) when the vendor capability block was omitted from the comparison mean.

Respondents who rated training distressed recorded a score of 2.74 ($n = 385$), against 5.09 ($n = 413$) among the rest (cluster-robust $p < 0.001$). Of the 385, 244 (63.38%) had a score of 3 or below. Among the 402 who rated training 4 to 7, 28 (6.97%) did, and none of the 11 who rated it 8 to 10 did (Monte Carlo chi-square = 284.31, $p < 0.001$, Cramér's V = 0.60).

The association was observed within the canceled engagements at $r = 0.72$ and ranged across the active-stage groups from 0.67 in Build-Out to 0.80 in Production. After Holm adjustment, it was significant in 85 of the 86 segment levels tested. Within role it stood at 0.80 among individual contributors, 0.77 among engagement managers, and 0.64 among portfolio directors. Among individual contributors alone, 65.59% (204 of 311) of those who rated training 1 to 3 had a score of 3 or below.

Five respondents on 5 engagements wrote about regulatory training, and all five answers were negative. None described one of the 26 forms of a control bypassed or not working. The answers described individual contributors placed on regulated work without training in its rules, who learned sanctions lists, suitability rules, and returns from colleagues or from the financial institution's own manuals.

> *"To be fair the hands on agent training was useful, I use it daily. The regulatory side they could not teach us, we learn that from the client's returns"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Regulatory, audit, tax and legal

The Fit Between Governance Processes and the Pace of Agent Work

Reported mismatches between control processes and the pace of agent work were associated with the governance, oversight, and control score in different ways. Forums and change-control cycles showed negative adjusted engagement relationships, while several other timing or capacity patterns did not. The common substantive concern is fit: authority, review, and evidence must operate on the work's timescale. The variation in results requires an account of each mechanism and its limits.

The written accounts describe five broad explanations: slow response by the firms, processes inherited from people-based work, limited experience with agentic AI, output volume and change frequency, and variable output with inadequate evidence. These are respondent explanations rather than identified causes. Each gathers distinct patterns. Their analysis compares respondent scores and relates the engagement share describing a pattern to the engagement score, before and after adjustment. Conditional source populations are stated separately.

Monthly governance forums and change frequency had negative adjusted engagement associations within the process-answer base (partial $r = -0.29$ and -0.38, $n = 133$). Their raw associations and conditional respondent contrasts were much less pronounced. Other accounts identified handoff, acceptance, review, documentation, and knowledge gaps, with different or nonsignificant score relationships. This variation limits a single statistical explanation based on pace. It also shows why each reported mechanism requires its own interpretation alongside the broader score.

Monthly forums, reliance on a person noticing errors, and excessive change frequency were classified within answers on manual alteration of processes. Their principal comparisons use the 516 respondents answering that question and the 133 engagements with an answer. Other patterns retain their specified written-answer sources. Table 3.14 reports each pattern's base, respondent comparison, and unadjusted and adjusted engagement association, including null findings.

Table 3.14. The written patterns under the five reasons against the score

Reason and written pattern	Base	Answers, respondents, engagements	The score, describing against the rest (p)
Slow Response From the Vendor and the Financial Institution			
Governance Forums Meeting Monthly Against Work Moving Daily	Answers on process manual alteration, 516 respondents	44, 41, 30	3.10 against 3.24 (p = 0.604)
Waiting as an Unmeasured Cost	All written answers	10, 10, 10	2.99 against 4.00 (p = 0.012)
Absence of an Escalation Path for Agent Failure Outside Working Hours	All written answers	14, 14, 14	3.91 against 3.99 (p = 0.845)
Governance Functions Inherited From Before Agentic AI			
Change, Risk, Access, and Quality Processes Designed for People	Answers on process manual alteration, 516 respondents	4, 4, 4	2.79 against 3.23 (p = 0.559)
Processes Dependent on a Person Noticing the Error	Answers on process manual alteration, 516 respondents	25, 24, 19	3.13 against 3.23 (p = 0.672)
Checks Retained in Individual Knowledge and Lost to Automation	All written answers	23, 22, 17	3.84 against 4.00 (p = 0.658)
Service Levels Unsuited to Agent Work	All written answers	23, 20, 15	4.76 against 3.97 (p = 0.107)
Documentation Trailing the Build	All written answers	46, 46, 38	4.93 against 3.94 (p < 0.001)
Undefined Handoffs Between the Financial Institution's Own Teams	All written answers	41, 39, 35	2.81 against 4.05 (p < 0.001)
The Newness of Agentic AI to Both Firms			
Practice Developed on the Live Engagement	All written answers	95, 93, 71	3.76 against 4.02 (p = 0.199)
Knowledge Transfer Not Completed	All written answers	25, 24, 22	3.64 against 4.00 (p = 0.241)
Unclear Liability and Regulatory Accountability	All written answers	83, 80, 57	4.32 against 3.96 (p = 0.134)

Table 3.14. The written patterns under the five reasons against the score (continued)

Reason and written pattern	Base	Answers, respondents, engagements	The score, describing against the rest (p)
Comparison With Prior Engagements	All written answers	237, 206, 117	4.35 against 3.88 (p = 0.002)
Volume and Speed of AI Output Beyond the Controls			
Human Review Overrun by Agent Volume	Negative answers on review capacity	123, 115, 80	4.13 against 3.97 (p = 0.409)
Change Frequency Beyond the Capacity of Change Control	Answers on process manual alteration, 516 respondents	101, 93, 65	3.13 against 3.25 (p = 0.477)
Non-Deterministic Output and Missing Logs			
Output That Cannot Be Signed Off	All written answers	29, 27, 23	2.34 against 4.05 (p < 0.001)
Absence of a Quality Assurance Method for Variable Output	Negative answers on quality assurance for variable output	126, 116, 81	3.60 against 4.05 (p = 0.010)
Missing Logs and Evidence of Agent Actions	Respondents rating both items	184 rated both 1 to 3	2.06 against 5.00 (neither)

Reason and written pattern	Engagement r (n)	Adjusted partial r (p)
Slow Response From the Vendor and the Financial Institution		
Governance Forums Meeting Monthly Against Work Moving Daily	−0.06 (133)	−0.29 (p < 0.001)
Waiting as an Unmeasured Cost	−0.09 (150)	−0.07 (p = 0.390)
Absence of an Escalation Path for Agent Failure Outside Working Hours	0.02 (150)	−0.02 (p = 0.787)
Governance Functions Inherited From Before Agentic AI		
Change, Risk, Access, and Quality Processes Designed for People	−0.09 (133)	−0.11 (p = 0.230)

Table 3.14. The written patterns under the five reasons against the score (continued)

Reason and written pattern	Base	Answers, respondents, engagements	The score, describing against the rest (p)
Processes Dependent on a Person Noticing the Error		−0.10 (133)	−0.07 (p = 0.400)
Checks Retained in Individual Knowledge and Lost to Automation		−0.04 (150)	−0.01 (p = 0.949)
Service Levels Unsuited to Agent Work		0.18 (150)	0.05 (p = 0.555)
Documentation Trailing the Build		0.25 (150)	−0.08 (p = 0.305)
Undefined Handoffs Between the Financial Institution's Own Teams		−0.28 (150)	0.01 (p = 0.876)
The Newness of Agentic AI to Both Firms			
Practice Developed on the Live Engagement		−0.12 (150)	−0.15 (p = 0.065)
Knowledge Transfer Not Completed		−0.10 (150)	0.09 (p = 0.250)
Unclear Liability and Regulatory Accountability		0.09 (150)	−0.06 (p = 0.485)
Comparison With Prior Engagements		0.24 (150)	0.00 (p = 0.974)
Volume and Speed of AI Output Beyond the Controls			
Human Review Overrun by Agent Volume		0.07 (150)	0.00 (p = 0.954)
Change Frequency Beyond the Capacity of Change Control		−0.08 (133)	−0.38 (p < 0.001)
Non-Deterministic Output and Missing Logs			
Output That Cannot Be Signed Off		−0.32 (150)	−0.11 (p = 0.180)
Absence of a Quality Assurance Method for Variable Output		−0.16 (150)	−0.11 (p = 0.178)
Missing Logs and Evidence of Agent Actions		0.94 (150)	Not tested

Note: Monthly forums, reliance on a person noticing errors, and change frequency use the 516 respondents answering on process alteration and 133 engagements with an answer for their principal comparisons. Any all-study or all-individual-contributor comparison is supplementary and uses a different population. Other patterns retain their stated written-answer sources and analysis bases. Negative-classified source answers are not an independent survey of every respondent. Written nonmention is not confirmed absence. Adjusted-correlation p values are nominal.

1 Slow Response From the Vendor and the Financial Institution

Decision and response timing connect formal governance with continuing agent work. Joint decision-speed ratings and financial institution response ratings were associated with the wider score, while monthly-forum accounts had a conditional engagement association. Waiting and after-hours escalation accounts supplied smaller and less conclusive evidence. These findings identify different points at which operational response may lag, rather than one verified cause shared by every engagement.

1.1 Governance Forums Meeting Monthly Against Work Moving Daily

Written pattern · Answers on the manual alteration of processes, read under the Part 5 classification of reasons · 44 answers from 41 respondents on 30 engagements

Governance forums that met on a calendar against daily work were associated with lower scores in the adjusted engagement comparison. The substantive concern is a decision cycle that may not match the pace of agent work. Within respondents who answered on process alteration, however, the mean-score contrast was not clearly detected. Those findings use different units and support a qualified account of timing mismatch.

Figure 3.81. Governance Forums Meeting Monthly Against Work Moving Daily: unadjusted and adjusted associations in the 133 engagements with an answer on process alteration.

Governance forums that met on a calendar against daily work were associated with lower scores in the adjusted engagement comparison.

Comparison	Estimate and analysis base
Unadjusted engagement r	r = −0.06 133 engagements
Adjusted engagement partial r	partial r = −0.29 133 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

Raw p = 0.484, adjusted nominal p < 0.001. The respondent comparison within 516 process-answer respondents is separate. These are conditional associations, not causal effects.

At the engagement level, within the 133 engagements with an answer on process manual alteration, the share describing the pattern was associated with the score at r = −0.06 (p = 0.484). After adjustment for the mean of the other 53 indicators, the association reached

partial r = −0.29 ($p < 0.001$). Across all 150 engagements the adjusted association stood at −0.32 and held between −0.33 and −0.27 when each category's block of the 53 was omitted from the comparison mean. The share describing the pattern was associated with the 53 indicators themselves at only r = −0.08, so the pattern marked weak governance on engagements whose other indicators did not.

The following distributions provide supplementary all-study context and do not replace the eligible process-answer comparisons. The pattern was described by 7.77% (22 of 283) of respondents with a score of 3 or below, 3.60% (19 of 528) above 3 to 7, and none of the 39 above 7. It was described by 6.80% of respondents in Transition/Cut-Over and 6.52% on canceled engagements, against 1.14% in Build-Out and 0.97% in Production. Individual contributors described it at 5.70% (38 of 667) and engagement managers at 2.00% (3 of 150), with no clearly detected difference by role after Holm adjustment. Among individual contributors alone, the 38 who described it stood at 3.18 against 3.91 among the other 629 (cluster-robust p = 0.009).

The 44 answers came from the answers on the manual alteration of processes, where the reading assigned each answer naming a forum, committee, governance body, panel, or approval loop as the reason for the manual alteration, and 23 of them were written on canceled engagements. Respondents described a sign-off forum that met monthly while the work ran daily, and a two-hour change that waited eleven days for a fortnightly forum.

> *“Decision rights are written down and that part is fine. The issue is timing. A sign-off on the lineage mapping goes to their data governance forum which meets fortnightly, so a two hour change waits eleven days”*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Technology engineering and operations

Part 5 (Manual Alteration of Financial Institution Processes, section 2.4) analyzes the same reading as a reason for process manual alteration.

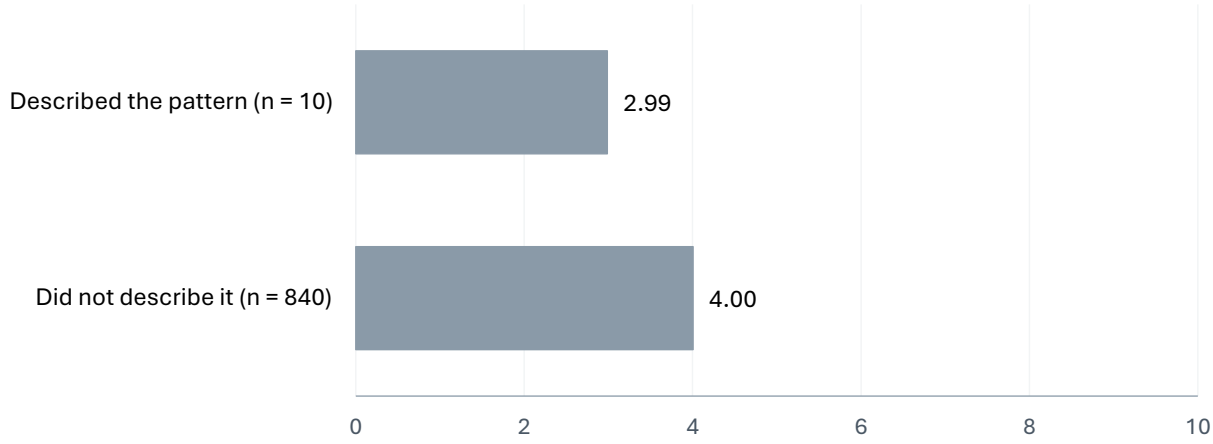
1.2 Waiting as an Unmeasured Cost

Written pattern · All written answers to the open questions · 10 answers from 10 respondents on 10 engagements

Respondents describing waiting as an unmeasured cost had lower mean scores, although the small pattern did not show a clear engagement-level association. A delay between recognizing a need and obtaining a decision or action can remain absent from performance measures. Ten accounts support a limited description of that issue. They do not establish the extent of waiting among other respondents.

Figure 3.82. Waiting as an Unmeasured Cost: descriptive mean score comparison.

Respondents describing waiting as an unmeasured cost had lower mean scores, although the small pattern did not show a clear engagement-level association.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

Mean scores were 2.99 for those describing waiting and 4.00 for the remainder. The cluster-robust test on the ten against the other 840 reached p = 0.012. Among individual contributors alone, the five who described waiting stood at 3.36 against 3.87 (cluster-robust p = 0.416).

At the engagement level the share describing waiting was associated with the score at r = −0.09 (p = 0.263, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.07 (p = 0.390).

Waiting was described by 7 of the 283 respondents with a score of 3 or below, 3 of the 528 above 3 to 7, and none of the 39 above 7. Five of the ten were engagement managers, and no segment family differed after Holm adjustment.

The ten answers came from all written answers to the open questions, and 4 of them were written on canceled engagements. Respondents described the days between a change by the financial institution and its arrival on the delivery team, and the weeks between a joint meeting and a decision that held, as costs that appeared in no measure.

> *"Nothing here asks about the waiting. Client change their conduct rule on a Monday, we hear about it maybe Thursday from a teammate, and everything flagged in between has to be reviewed again by hand. That part never appear anywhere."*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Regulatory, audit, tax and legal

> *"The joint review side actually worked. Two of their buyers sat with us on the Tuesday call and cleared the flagged POs same day. What no box asks about is the three weeks between one of those calls and any decision that stuck."*
>
> **ENGAGEMENT MANAGER**
> CXM, Corporate and back-office functions

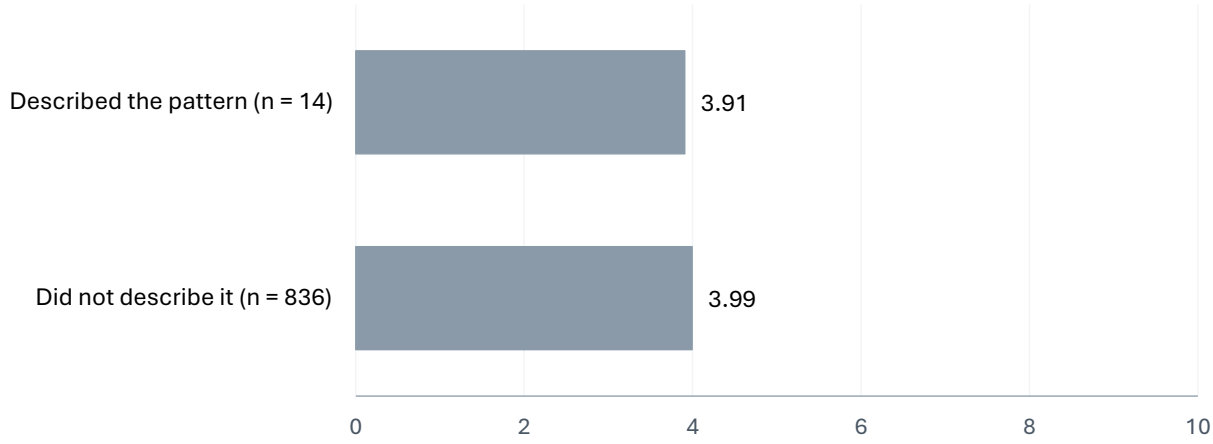
1.3 Absence of an Escalation Path for Agent Failure Outside Working Hours

Written pattern · All written answers to the open questions · 14 answers from 14 respondents on 14 engagements

Accounts of missing after-hours escalation paths did not show a clearly detected score relationship. The operational concern remains specific: agent work may continue when the people authorized to respond are unavailable. A broad score may coexist with this reported gap at different levels. The 14 written descriptions support retaining the issue while limiting any conclusion about prevalence or comparative control performance.

Figure 3.83. Absence of an Escalation Path for Agent Failure Outside Working Hours: descriptive mean score comparison.

Accounts of missing after-hours escalation paths did not show a clearly detected score relationship.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 14 respondents who described it, a small base, stood at 3.91 and the other 836 at 3.99 (cluster-robust p = 0.845). Among individual contributors alone, the 6 who described it stood at 3.85 against 3.86 (cluster-robust p = 0.976).

At the engagement level the share describing it was associated with the score at r = 0.02 (p = 0.808, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.02 (p = 0.787).

The pattern was described by 5 of the 283 respondents with a score of 3 or below, 7 of the 528 above 3 to 7, and 2 of the 39 above 7, a descriptive comparison that does not establish equal prevalence. It was described by 6 individual contributors, 6 engagement managers, and 2

portfolio directors, with no clearly detected difference by role after Holm adjustment, and by 1.82% of respondents in the lowest indicator quintile against 1.88% in the highest.

The 14 answers came from all written answers to the open questions, and 7 of them were written on canceled engagements. Respondents described agents that failed at two in the morning with no person rostered to decide, and a portfolio director on a higher-scoring engagement described the hours each day when only one of the two firms was at work.

> *“Nothing here asks who we call at 2am when the agent starts closing tickets it shouldn't. There is no such person at our firm, we work it out ourselves.”*
>
> **ENGAGEMENT MANAGER**
>
> CXM, Corporate and back-office functions

> *“The sixes in this section are not a general complaint about how we run the thing together, they come from one narrow gap. Governance in the daytime is genuinely good, and I would say our joint quality loop catches almost everything before it reaches a live endpoint. But the bank's change desk keeps office hours and our desk does not. So the questions about stopping something mid-flight, or holding one view of the risk, are really questions about the eleven hours a day when only one of the two firms is awake. And nobody has written down who makes that call yet. We have not been caught by it, which is luck as much as design.”*
>
> **PORTFOLIO DIRECTOR**
>
> ITO, Technology engineering and operations

“

there was a case last month where an escalation sat for two days, their process assumes a human will notice and the agent does not raise flags the same way. Moreover their governance model assumes changes happen at release pace, not the pace an agent generates cases, so we adapted the workflow ourselves.”

INDIVIDUAL CONTRIBUTOR

KPO, Financial crime and client onboarding

2 Governance Functions Inherited From Before Agentic AI

Processes inherited from people-based delivery were described as poorly fitted to agent work in several distinct ways. Some depended on a person noticing an error, some left knowledge unwritten, and others used service levels, documentation, or handoffs that respondents considered inadequate. Their score relationships varied. Together, these findings locate the shared governance concern in whether controls have been adapted to the work agents perform.

2.1 Change, Risk, Access, and Quality Processes Designed for People

Written pattern · Answers on the manual alteration of processes, read under the Part 5 classification of reasons · 4 answers from 4 respondents on 4 engagements

Four accounts described processes designed for people being carried into agent work without sufficient adaptation. Their governance relevance lies in the fit between operational authority, review, and the pace of work. The evidence is descriptive and too small for the paper's comparative tests. The separately rated joint-model indicator provides related evidence about that fit. One account in each status group does not establish a percentage gradient.

Figure 3.84. Change, Risk, Access, and Quality Processes Designed for People: descriptive evidence and its limits

Four accounts described processes designed for people being carried into agent work without sufficient adaptation.

Reported practice	Governance implication	Evidence boundary
Four accounts described processes designed for people being carried into agent work without sufficient adaptation.	Their governance relevance lies in the fit between operational authority, review, and the pace of work.	The evidence is descriptive and too small for the paper's comparative tests.

Note: Respondents describing change, Risk, Access, and Quality Processes Designed for People

Small base. No percentage comparison between status groups.

4 written accounts from 4 respondents on 4 engagements.

Mean score 2.79 among those describing the pattern, compared with 4.00 among the remainder. Descriptive comparison only, with no inferential test.

The four respondents recorded a score of 2.79 against 4.00 among the rest, a contrast printed as description only. Three of the four stood with a score of 3 or below and one above 3 to 7, and

the four answers came one each from Build-Out, Transition/Cut-Over, Production, and a canceled engagement.

The four accounts were classified within answers on manual alteration of processes. They described an audit workflow unchanged from people-based work and processes written before agents were introduced. The separately rated fit of the joint model to agent work had a positive adjusted cross-category association (partial r = 0.26). That rating supplies related evidence, with no inference that it verifies the mechanism proposed in four accounts.

> *This is not one time thing, happen every case here... I tell my lead every time but nothing come back, we just carry on. The audit workflow they gave us was written for people, not for agent, and nobody on their side has looked at it since before the engagement started. So what we do is rebuild the steps as we go and keep own notes. It works, cases close, but the documented process and the actual process are two different things now and everybody knows it and nobody says it out loud. I have kindly asked three times for someone from their side to sit with us for one hour and walk the exceptions. Three times. Nothing. Meanwhile the volumes go up and the sign off window stays the same. My lead is fine, he raise it upward as far as i can tell, but it goes into whatever meeting decides these things and does not come out again. And the frustrating part is when something does go wrong it comes back to us, not to the process, not to the people who wrote it. It comes back to whoever touched it last which is"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Regulatory, audit, tax and legal

> *But kindly note this is not a data problem, it is that the process was written before the agent existed. I flag the same issue every review and nothing changes"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Regulatory, audit, tax and legal

Part 5 (Manual Alteration of Financial Institution Processes, section 2.4) analyzes the same reading as a reason for process manual alteration.

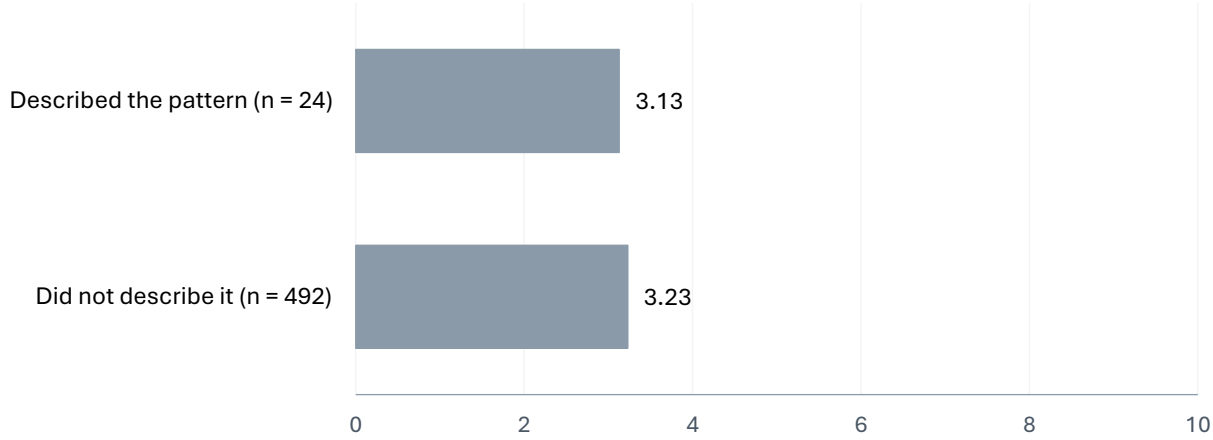
2.2 Processes Dependent on a Person Noticing the Error

Written pattern · Answers on the manual alteration of processes, read under the Part 5 classification of reasons · 25 answers from 24 respondents on 19 engagements

Dependence on a person noticing an error was reported within process-alteration accounts without a clearly detected conditional mean-score contrast. The control concern is detection that relies on individual attention rather than an explicit and consistently operating check. The pattern is distinct from automated monitoring and from knowledge that never entered the process. The principal comparison uses eligible process answers, with all-study comparisons presented separately as context.

Figure 3.85. Processes Dependent on a Person Noticing the Error: descriptive mean score comparison.

Dependence on a person noticing an error was reported within process-alteration accounts without a clearly detected conditional mean-score contrast.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The base is respondents answering on process alteration.

The contrast within the 516 respondents who wrote an answer on process manual alteration reached cluster-robust p = 0.672. Against all 850 respondents it widened to 3.13 against 4.02. This wider comparison also changes the population, so it does not isolate the contribution of the specific written pattern.

Within the 133 engagements with an answer on process manual alteration, the share describing the pattern was associated with the score at r = −0.10 (p = 0.250), and after adjustment for the mean of the other 53 indicators at partial r = −0.07 (p = 0.400).

The following distributions provide supplementary all-study context and do not replace the eligible process-answer comparisons. The pattern was described by 4.59% (13 of 283) of

respondents with a score of 3 or below, 2.08% (11 of 528) above 3 to 7, and none of the 39 above 7, and by 3.03% in the lowest indicator quintile against none in the highest. Twenty-three of the 24 respondents who described it were individual contributors. Among individual contributors alone, the 23 stood at 3.18 against 3.89 among the other 644 (cluster-robust p = 0.009), a contrast drawn against all individual contributors rather than within the base of the reading.

The 25 answers came from the answers on the manual alteration of processes, where the reading assigned each answer naming a check that depended on a person noticing, catching, or spotting the error, and 13 of them were written on canceled engagements. Respondents described workflows documented for an analyst who would notice a comparable that looked wrong, and individual contributors who built manual checks of their own to cover the step the agent skipped.

> *“their process leans on a person catching the miss, the agent has no such instinct, so i step in myself before the model runs. that is the whole thing really. the workflow they documented assumes an analyst who notices when a comparable looks wrong, and the agent does not notice, it just uses it. so somebody has to be the analyst. that is me most days. i have said in two different forums now that the process needs rewriting for how the tool actually behaves rather than how a person behaves, and both times people agreed and nothing followed. i think the difficulty is that rewriting it means somebody admits the original design was wrong, and the original design has names on it. so instead we have a documented process and a real process and i am the bridge between them. i am not unhappy about it, the work is interesting and i am learning a lot, in all honesty. i just think if i went on leave for two weeks somebody would find out how much of this was being held together by one person noticing things, and”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Corporate and back-office functions

> *what they did is build a process that expects a human to catch every mistake before it goes anywhere, but agent doesnt catch anything on its own, it just runs and moves to next case, so i built manual checks in myself to cover for that gap, please note this is not something anyone asked me to do, i just did it because gap was there and somebody had to. their systems never agree with each other on same figures, one say one thing other say another, and nobody ever opened up their api for us properly so we pull flat exports and patch from there instead, which take hours some days, and every access request just goes unanswered for weeks, i dont even chase it anymore because whats the point, i just work around same problem every single time it comes up. if I am honest this whole engagement is closing down soon anyway so i dont know why i am even explaining all this, but i wanted to say it properly at least once, because most days it feels like nobody upstream understands how much of this is held together by me checking things twice that agent should have caught the first time, and i dont discuss about it much with anyone because there isnt really anyone left to discuss about it with at this point, so i just"*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Technology engineering and operations

Part 5 (Manual Alteration of Financial Institution Processes, section 2.4) analyzes the same reading as a reason for process manual alteration.

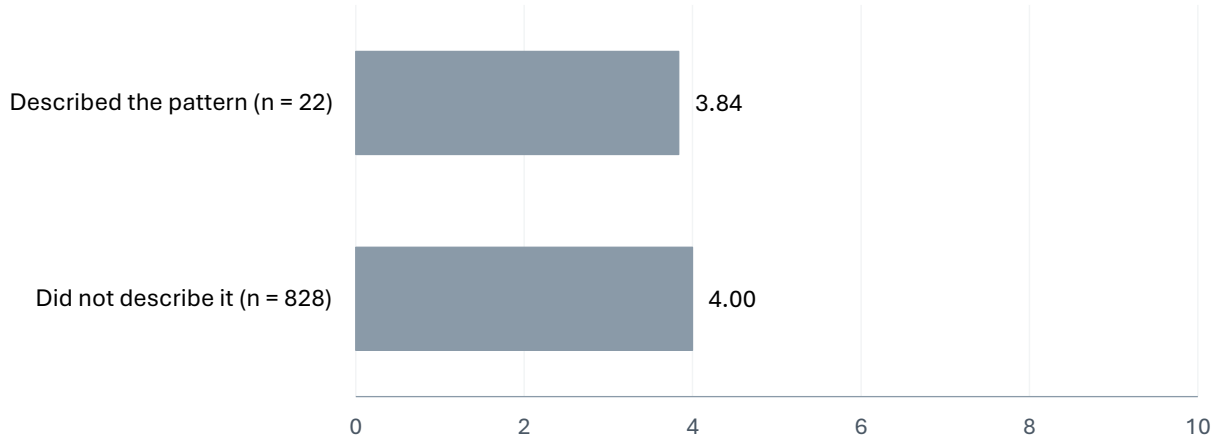
2.3 Checks Retained in Individual Knowledge and Lost to Automation

Written pattern · All written answers to the open questions · 23 answers from 22 respondents on 17 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 4.5

Accounts of checks retained in individual knowledge did not show a clearly detected score relationship. These descriptions identify a governance concern that can be lost when work is automated without making its informal checks explicit. The small written base supports an account of the mechanism respondents described, while the study does not observe the transfer of each check or establish its causal contribution to an outcome.

Figure 3.86. Checks Retained in Individual Knowledge and Lost to Automation: descriptive mean score comparison.

Accounts of checks retained in individual knowledge did not show a clearly detected score relationship.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The overall respondent contrast had p = 0.658. The 22 respondents who described the pattern, a small base, stood at 3.84 and the other 828 at 4.00. Among individual contributors alone, the 18 who described it stood at 3.94 against 3.86 among the other 649 (cluster-robust p = 0.859).

At the engagement level the share describing the pattern was associated with the score at r = −0.04 (p = 0.643, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.01 (p = 0.949).

The pattern was described by 3.18% (9 of 283) of respondents with a score of 3 or below, 2.27% (12 of 528) above 3 to 7, and 1 of the 39 above 7. It was described by 18 individual contributors, 2 engagement managers, and 2 portfolio directors, with no clearly detected

difference by role after Holm adjustment, and no segment family differed after Holm adjustment.

The 23 answers came from all written answers to the open questions, and 11 of them were written on canceled engagements. Part 5 counts 14 answers on 9 engagements for the same pattern, among the written answers on the manual alteration methods alone, so the two counts rest on different bases. Respondents described a check carried in one person's head and spreadsheet that the agent did not perform and that would leave with that person, and a process that assumed a person would catch the edge cases the agent passed over.

> *"I patch the Vendor records by hand every single day just to keep the Desk running, I told my manager and hes aware, thats not the issue. The issue is the master data on their side has three sources and they disagree with each other, so depending which one the extract pulls from you get a different supplier address, and the agent cannot tell which is right. I can. Usually. Because Ive seen these vendors for eight months and I know which ones moved office and which ones have two entities. That knowledge is entirely in my head and in a spreadsheet on my machine. Nobody has ever asked me to write it down. If I left tomorrow the Desk would start pushing purchase orders to old addresses within a week and nobody would know why. Ive mentioned that. People nod. Its the sort of thing everyone agrees is a problem and nobody has time to fix because fixing it means somebody has to reconcile three master data sources at a bank, which is a project, not a task. So we carry on. And in truth the work itself is fine, its steady, I just think"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Corporate and back-office functions

> *"I adjusted the triage step because their process assumes a human catches the edge cases and the agent just doesn't, so I added a manual check before it goes further"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

Part 5 (Leading Dynamics of Manual Alteration, section 4.5) analyzes the pattern as a factor in manual alteration.

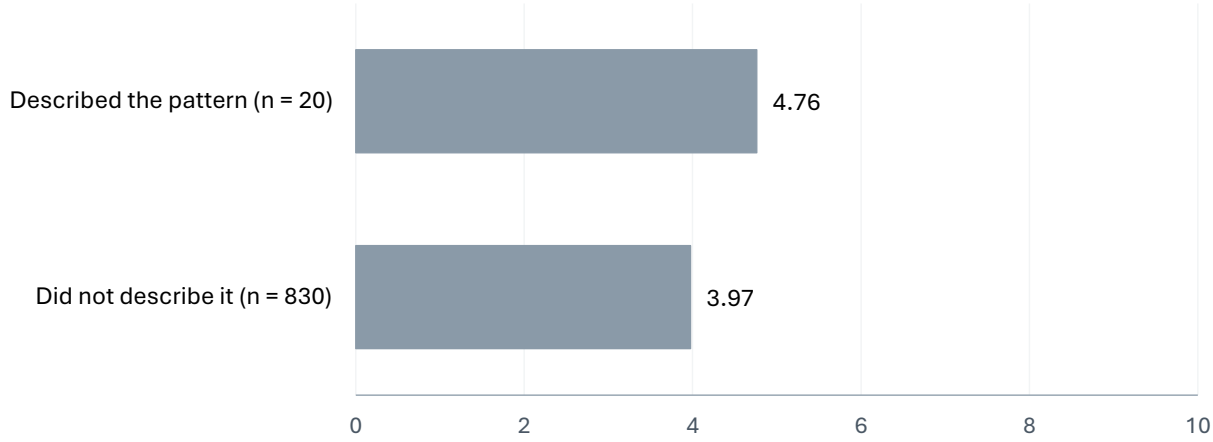
2.4 Service Levels Unsuited to Agent Work

Written pattern · All written answers to the open questions · 23 answers from 20 respondents on 15 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 8.4

Service levels unsuited to agent work were described on engagements with somewhat higher observed scores, and the adjusted relationship was not clearly detected. This qualifies a simple reading in which every reported deficiency must accompany a lower composite. The governance question is whether the service commitment measures the pace and form of work now performed. The positive observed association does not establish that the mismatch improves control.

Figure 3.87. Service Levels Unsuited to Agent Work: descriptive mean score comparison.

Service levels unsuited to agent work were described on engagements with somewhat higher observed scores, and the adjusted relationship was not clearly detected.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 20 respondents who described the pattern, a small base, recorded a score of 4.76, against 3.97 among the other 830 (cluster-robust p = 0.107). Among individual contributors alone, the 15 who described it stood at 4.67 against 3.85 (cluster-robust p = 0.140).

At the engagement level the share describing the pattern was associated with the score at r = 0.18 (p = 0.028, n = 150 engagements). After adjustment for the mean of the other 53 indicators the partial r fell to 0.05 (p = 0.555). The adjusted comparison did not detect a clear association beyond the comparison mean.

The pattern was described by 5.00% of respondents in the highest indicator quintile against 0.61% in the lowest, by 3.88% in Production against 1.55% on canceled engagements, and by 1.41% (4 of 283) of respondents with a score of 3 or below against 2.84% (15 of 528) above 3 to 7. No role or segment family differed after Holm adjustment.

The 23 answers came from all written answers to the open questions, and 7 of them were written on canceled engagements. Part 5 counts 4 answers on 4 engagements for the same pattern, among the written answers on the manual alteration methods alone, so the two counts rest on different bases. Respondents on active engagements described service level and quality targets written for human pace that lost their meaning once agents handled the work.

Part 5 (Leading Dynamics of Manual Alteration, section 8.4) analyzes the pattern as a factor in manual alteration.

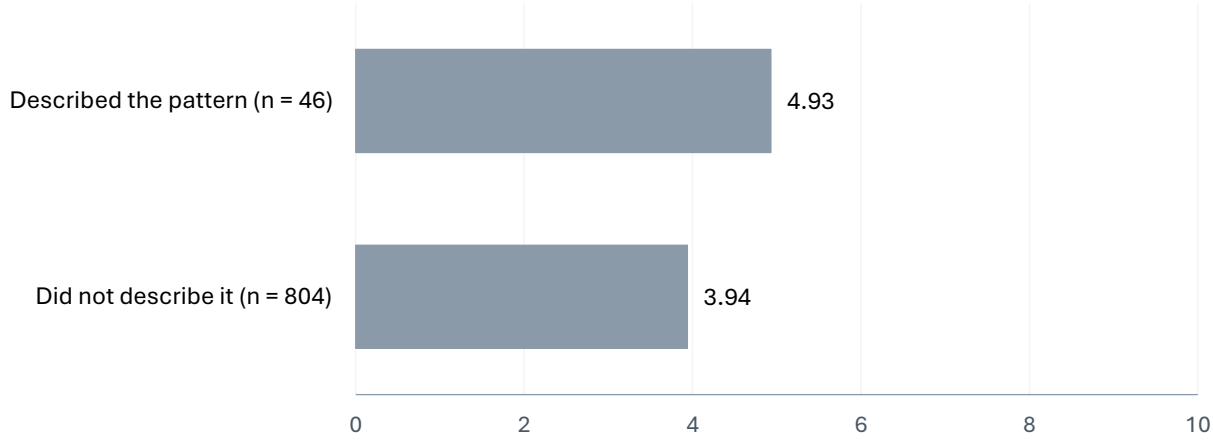
2.5 Documentation Trailing the Build

Written pattern · All written answers to the open questions · 46 answers from 46 respondents on 38 engagements · Listed in Part 5, Additional Conditions Below the Threshold for Statistical Analysis

Documentation lag was described by respondents with higher mean scores, while the adjusted association was not clearly detected. This finding distinguishes reported awareness of a gap from the overall severity of governance weakness. The operational question is whether the record used for oversight reflects the system currently in use. The comparison does not show that better governance causes documentation to lag or that lag is harmless.

Figure 3.88. Documentation Trailing the Build: descriptive mean score comparison.

Documentation lag was described by respondents with higher mean scores, while the adjusted association was not clearly detected.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 46 respondents who described it recorded a score of 4.93, against 3.94 among the other 804 (cluster-robust $p < 0.001$). Among individual contributors alone, the 33 who described it stood at 4.97 against 3.81 (cluster-robust $p < 0.001$).

At the engagement level the share describing the pattern was associated with the score at $r = 0.25$ ($p = 0.002$, $n = 150$ engagements). After adjustment for the mean of the other 53 indicators the partial r was −0.08 ($p = 0.305$). The adjusted comparison did not detect a clear association beyond the comparison mean.

The pattern was described by 11.65% of respondents in Production against 2.48% on canceled engagements, and by 9.38% in the highest indicator quintile against 1.82% in the lowest. By the

score it was described by 2.12% (6 of 283) of respondents at 3 or below, 6.82% (36 of 528) above 3 to 7, and 4 of the 39 above 7. The share differed after Holm adjustment by class of work and by no other segment family.

The 46 answers came from all written answers to the open questions, and 8 of them were written on canceled engagements. Part 5 lists 2 answers on 2 engagements for the same pattern below its threshold for statistical analysis, among the written answers on the manual alteration methods alone. Documentation that trailed the build was a complaint of engagements that had built enough to document, including the highest-scoring engagement of the study, where an individual contributor named it among the three challenges of the work.

> *“Last month a servicing rule changed mid week and process still assumed one change a month, besides there was no time allowed to update documentation properly, so exception just get handled by whoever pick it up”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Payments, cards and disputes

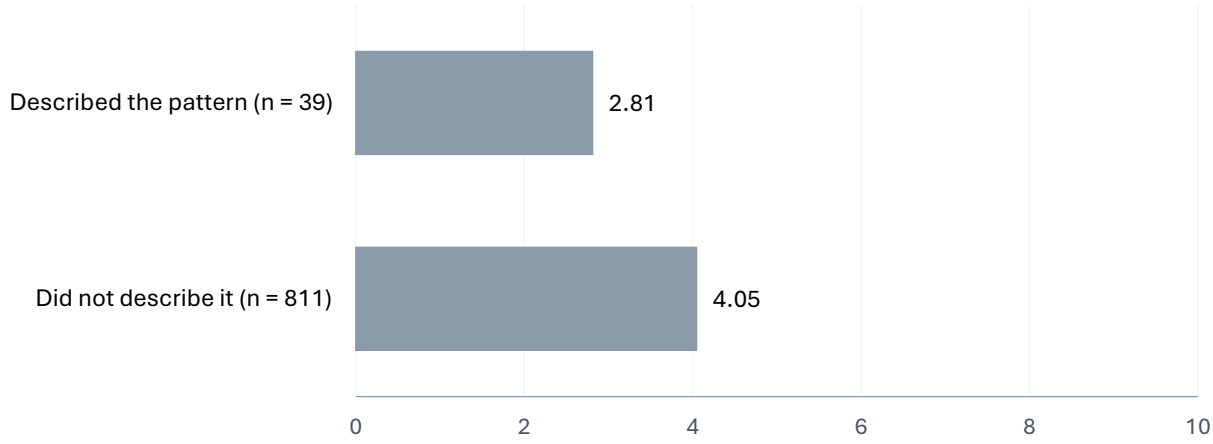
2.6 Undefined Handoffs Between the Financial Institution's Own Teams

Written pattern · All written answers to the open questions · 41 answers from 39 respondents on 35 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 3.5

Undefined handoffs within the financial institution were described by respondents with lower mean scores. The control concern is responsibility at the boundary between the institution's own teams, which can remain unresolved even where the vendor's task is clear. The adjusted engagement comparison did not detect a separate association. Lack of mention in the highest quintile records no observed descriptions there, rather than verified absence of the problem.

Figure 3.89. Undefined Handoffs Between the Financial Institution's Own Teams: descriptive mean score comparison.

Undefined handoffs within the financial institution were described by respondents with lower mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 39 respondents who described them recorded a score of 2.81, against 4.05 among the other 811 (cluster-robust $p < 0.001$). Among individual contributors alone, the 34 who described them stood at 2.89 against 3.92 (cluster-robust $p < 0.001$).

At the engagement level the share describing the pattern was associated with the score at $r = -0.28$ ($p < 0.001$, $n = 150$ engagements). After adjustment for the mean of the other 53 indicators the partial r was 0.01 ($p = 0.876$). The adjusted comparison did not detect a clear association beyond the comparison mean.

The pattern was described by 8.13% (23 of 283) of respondents with a score of 3 or below, 3.03% (16 of 528) above 3 to 7, and none of the 39 above 7. It was described by 7.45% of

respondents on canceled engagements against 1.94% to 4.00% across the active stages, and by 7.88% in the lowest indicator quintile against none in the highest. The share differed after Holm adjustment by the respondent's understanding of the governing obligations and by regulatory training.

The 41 answers came from all written answers to the open questions, and 25 of them were written on canceled engagements. Part 5 counts 29 answers from 27 respondents on 24 engagements for the same pattern, among the written answers on the manual alteration methods alone, so the two counts rest on different bases. Respondents described requirements from two of the financial institution's offices that conflicted with no rule for which prevailed, and discrepancies that had no owner to whom they could be taken.

> *If you read one thing here, read this. Nobody on their side owned the surveillance rules. We would get a requirement from compliance in London and a different one from the New York desk for the same scenario, and neither would say which one wins. They answer email fast, Ill give them that, but a fast answer from someone with no authority doesnt move anything. Their lead on this changed twice in eight weeks, so we explained the same design twice to two different people."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Regulatory, audit, tax and legal

> *There is nothing here that the numbers do not already say. I will put it down anyway. When a discrepancy comes out of the anomaly run there is no person to take it to. I asked in the second month, my lead asked also, and the answer was that it depends on the item. That is not an answer. So the practice is, whoever is on the queue that day decides, we write a note in the tracker and move on. Moreover the client side has three teams in the chain and each one says the sign off is with the other, so nothing moves. I have sat in the joint call. It is a status call, nobody makes decision in it. The same items are read out week after week, the same items, and after a while we stopped putting the hard ones on the list because there is no point. On the quality side, both firms count how many exceptions we cleared. Nobody counts how many we cleared wrongly. That is the measure that is missing and I have said this before to my lead. Two firms and neither one owns the number that matters. I do not think anybody will read this box. The engagement is finished anyway..."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Regulatory, audit, tax and legal

Part 5 (Leading Dynamics of Manual Alteration, section 3.5) analyzes the pattern as a factor in manual alteration.

3 The Newness of Agentic AI to Both Firms

Learning, knowledge transfer, and accountability were described as unresolved features of agentic delivery across a range of scores. These patterns concern preparation for unfamiliar work and how experience is brought into the operating model. Their inconsistent statistical relationships limit a single explanation based on newness. Each pattern nonetheless identifies a distinct responsibility relevant to oversight.

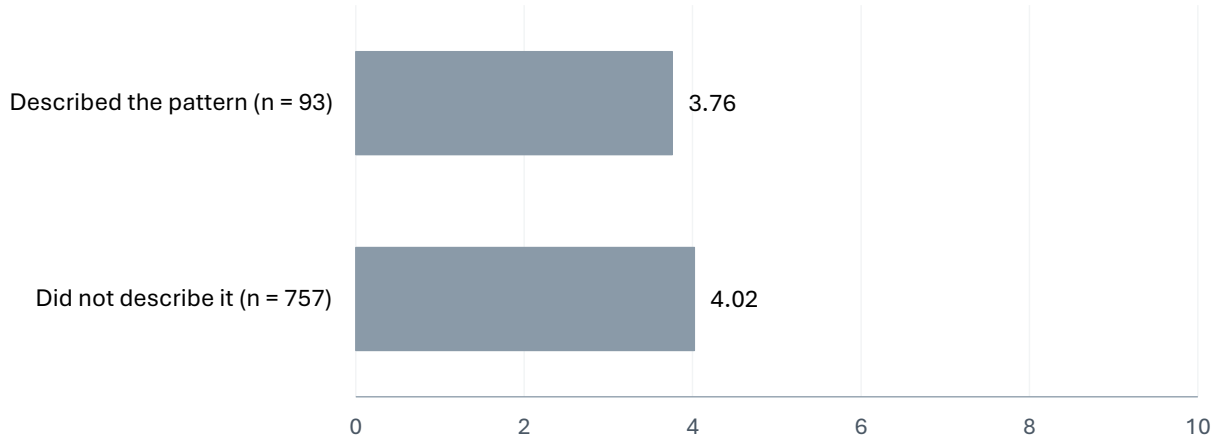
3.1 Practice Developed on the Live Engagement

Written pattern · All written answers to the open questions · 95 answers from 93 respondents on 71 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 4.3

Developing practice during live delivery was described across a range of scores without a clearly detected contrast. This identifies uncertainty in how the two firms learn to govern unfamiliar work. The governance question is whether authority, checks, and evidence develop alongside live practice or are settled beforehand. The written accounts describe that learning context, without measuring the effectiveness or causal consequences of learning on the engagement.

Figure 3.90. Practice Developed on the Live Engagement: descriptive mean score comparison.

Developing practice during live delivery was described across a range of scores without a clearly detected contrast.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The overall respondent contrast had p = 0.199. The 93 respondents who described it stood at 3.76 and the other 757 at 4.02. Among individual contributors alone, the 67 who described it stood at 3.54 against 3.90 (cluster-robust p = 0.082).

At the engagement level the share describing the pattern was associated with the score at r = −0.12 (p = 0.154, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.15 (p = 0.065).

The pattern was described by 13.07% (37 of 283) of respondents with a score of 3 or below, 10.04% (53 of 528) above 3 to 7, and 3 of the 39 above 7. It was described by 8.74% to 10.80% of respondents across the active stages and 12.73% on canceled engagements, and by 10.04% of individual contributors, 14.67% of engagement managers, and 4 of the 31 portfolio directors, with no clearly detected difference by role after Holm adjustment. No segment family differed after Holm adjustment.

The 95 answers came from all written answers to the open questions, and 42 of them were written on canceled engagements. Part 5 counts 12 answers from 11 respondents on 10 engagements for the same pattern, among the written answers on the manual alteration methods alone, so the two counts rest on different bases. Respondents described delivery teams that had never assured agent output at volume and invented their own checking, and teams that learned to run agents on the financial institution's live work.

> *“The playbook is fine on paper. We were never shown how to assure agent output at this volume, so we invented our own checking. Pricing is not solved either, whatever the deck says... Leadership is far from the floor. That is all.”*
>
> **ENGAGEMENT MANAGER**
>
> BPO, Trade, settlement and reconciliation

> *“These ones are for my desk, not the firm. Nobody here had run agents live before us, we learned on the client's tax memos. Elsewhere i am told it is better.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Regulatory, audit, tax and legal

Part 5 (Leading Dynamics of Manual Alteration, section 4.3) analyzes the pattern as a factor in manual alteration.

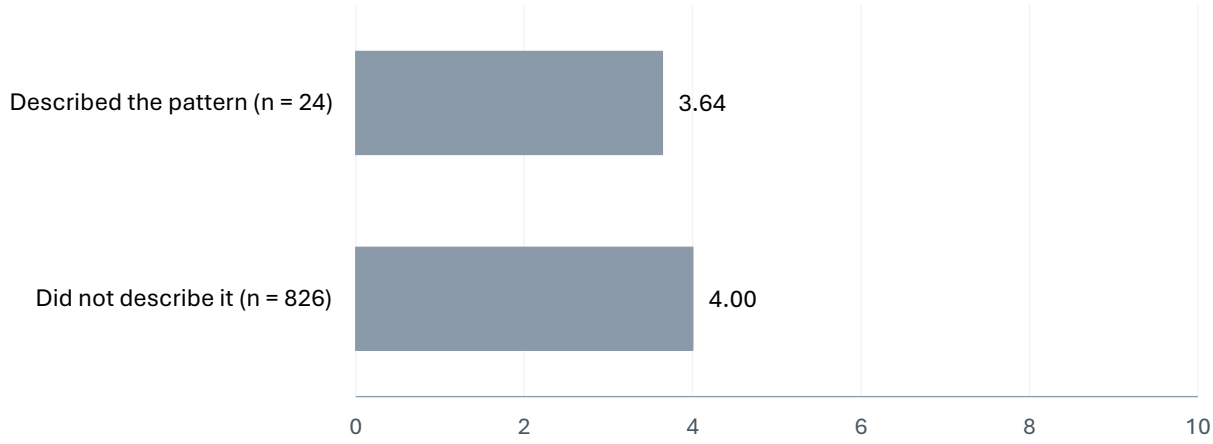
3.2 Knowledge Transfer Not Completed

Written pattern · All written answers to the open questions · 25 answers from 24 respondents on 22 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 4.4

Incomplete knowledge transfer was described without a clearly detected score relationship. Its relevance to control is the transfer of the operational knowledge needed to recognize exceptions, interpret output, and respond to problems. The small written base supports keeping the issue distinct from general experience or training. The result does not establish that knowledge transfer was equally complete across the remaining engagements.

Figure 3.91. Knowledge Transfer Not Completed: descriptive mean score comparison.

Incomplete knowledge transfer was described without a clearly detected score relationship.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The overall respondent contrast had p = 0.241. The 24 respondents who described it, a small base, stood at 3.64 and the other 826 at 4.00. Among individual contributors alone, the 21 who described it stood at 3.71 against 3.87 (cluster-robust p = 0.637).

At the engagement level the share describing the pattern was associated with the score at r = −0.10 (p = 0.202, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = 0.09 (p = 0.250).

The pattern was described by 3.53% (10 of 283) of respondents with a score of 3 or below, 2.65% (14 of 528) above 3 to 7, and none of the 39 above 7, and by 4.24% in the lowest

indicator quintile against 0.62% in the highest. No role or segment family differed after Holm adjustment.

The 25 answers came from all written answers to the open questions, and 12 of them were written on canceled engagements. Part 5 counts 6 answers on 5 engagements for the same pattern, among the written answers on the manual alteration methods alone, so the two counts rest on different bases. Respondents described knowledge transfer measured by the number of sessions held rather than by what was captured, and manual fixes by a former team that had never been written down.

> *"SME knowledge-transfer completion was tracked by the number of sessions held rather than whether the underwriting logic was actually captured, so the metric looked complete while real institutional knowledge kept leaving with departing staff."*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Payments, cards and disputes

> *"Nobody wrote down which bits the old team fixed by hand, we are rediscovering it case by case"*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Client servicing and communications

Part 5 (Leading Dynamics of Manual Alteration, section 4.4) analyzes the pattern as a factor in manual alteration.

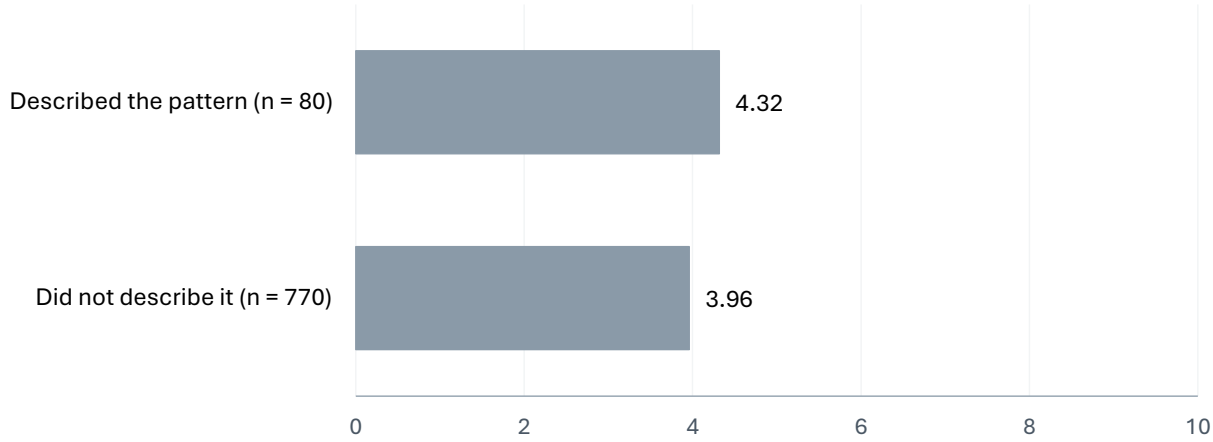
3.3 Unclear Liability and Regulatory Accountability

Written pattern · All written answers to the open questions · 83 answers from 80 respondents on 57 engagements · Listed in Part 5, Additional Conditions Below the Threshold for Statistical Analysis

Unclear liability and regulatory accountability were described across a range of scores. These accounts concern who is responsible when agent work causes an error, a different question from routine workflow ownership or a respondent's knowledge of governing obligations. The descriptions identify uncertainty rather than adjudicate liability. Their concentration in Production is a cross-sectional observation, with no inference about when accountability became unclear.

Figure 3.92. Unclear Liability and Regulatory Accountability: descriptive mean score comparison.

Unclear liability and regulatory accountability were described across a range of scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 80 respondents who raised it recorded a score of 4.32, against 3.96 among the other 770 (cluster-robust p = 0.134). Among individual contributors alone, the 58 who raised it stood at 4.21 against 3.83 (cluster-robust p = 0.138).

At the engagement level the share raising it was associated with the score at r = 0.09 (p = 0.275, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.06 (p = 0.485).

Unclear liability was raised by 9 of the 39 respondents with the score above 7 (23.08%), against 8.13% (23 of 283) at 3 or below and 9.09% (48 of 528) above 3 to 7. It was raised by 14.56% of

respondents in Production, against 9.14% to 9.60% in the earlier stages and 7.76% on canceled engagements, and by 11.25% in the highest indicator quintile against 9.70% in the lowest. The share differed after Holm adjustment by class of work.

The 83 answers came from all written answers to the open questions, and 26 of them were written on canceled engagements. Part 5 lists 2 answers on 2 engagements for the same pattern below its threshold for statistical analysis, among the written answers on the manual alteration methods alone. The question of who answered for an agent's error reached the highest-scoring engagements and those in Production. An individual contributor whose own score stood above 7, and who credited the engagement's staffing and training, was not sure who would answer for a costly agent error.

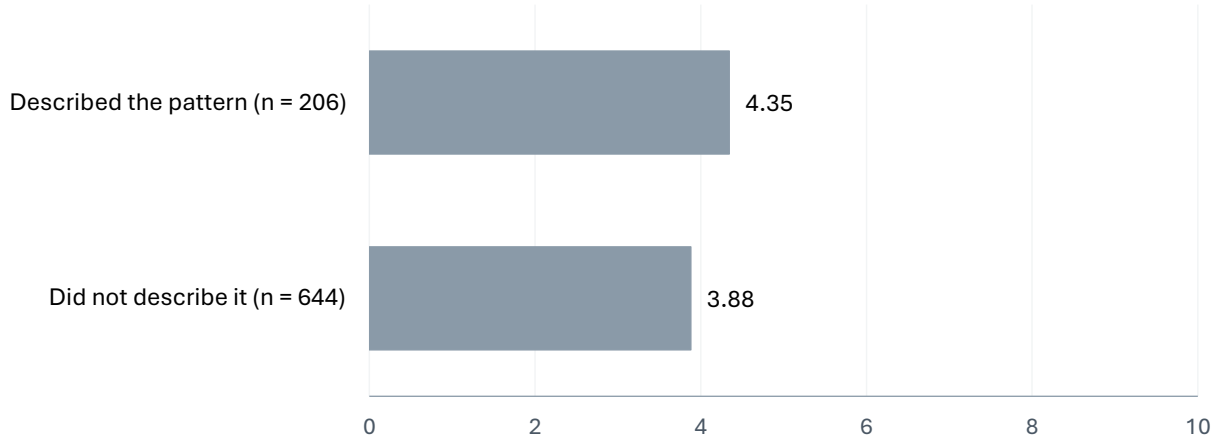
3.4 Comparison With Prior Engagements

Written pattern · All written answers to the open questions · 237 answers from 206 respondents on 117 engagements · Listed in Part 5, Additional Conditions Below the Threshold for Statistical Analysis

Respondents comparing the engagement with prior work had higher observed mean scores, while the adjusted association was not clearly detected. Their comparisons provide context for how they assess current control arrangements. They do not create a measured before-and-after study. The accounts describe respondents' frames of reference and do not establish that prior experience improved the present engagement's controls.

Figure 3.93. Comparison With Prior Engagements: descriptive mean score comparison.

Respondents comparing the engagement with prior work had higher observed mean scores, while the adjusted association was not clearly detected.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 206 respondents who made a comparison recorded a score of 4.35, against 3.88 among the other 644 (cluster-robust p = 0.002). Among individual contributors alone, the 142 who made one stood at 4.09 against 3.80 (cluster-robust p = 0.108).

At the engagement level the share making a comparison was associated with the score at r = 0.24 (p = 0.003, n = 150 engagements). After adjustment for the mean of the other 53 indicators the partial r was 0.00 (p = 0.974).

Comparison was made by 36.88% of respondents in the highest indicator quintile against 21.82% in the lowest, and by 38.46% (15 of 39) of respondents with the score above 7 against 21.91% (62 of 283) at 3 or below. It differed by role after Holm adjustment, made by 24 of the 31 portfolio directors (77.42%) against 26.67% of engagement managers and 21.29% of individual contributors, and portfolio directors recorded the highest the score of the three roles. It also differed by the manual alteration methods used and by disclosure to the financial institution.

The 237 answers came from all written answers to the open questions, and 89 of them were written on canceled engagements. Part 5 lists 1 answer on 1 engagement for the same pattern below its threshold for statistical analysis. The comparisons ran in both directions. On the highest-scoring engagement of the study, an individual contributor compared a change agreed and written down the same day with a weekly call on an earlier account, while on a canceled engagement an individual contributor recalled a previous project that had a playbook.

> *“On my last engagement any change had to wait for a weekly call, here the two sides agree a change same day and it is written down properly. Issues still go into two seperate trackers and we chase both, that part is not solved. Overall it runs better than what I saw before”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Credit, underwriting and suitability

> *“Felt behind from day one. My last project had an actual playbook. Here you learn it by watching the person next to you, half of them are guessing too. But nobody above us has shipped one of these before, so i don't know who'd write it.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Client servicing and communications

“

The part that reaches me is volume of review. Agents produce more test scripts in a morning than we can properly look at, so my day becomes reading output instead of building anything. i dont mind it exactly, team is good and my lead shields us where he can. Still i finish the week and cant say what i actually built. That is new for me, in addition nobody planned for it, it just arrived with the tooling.”

INDIVIDUAL CONTRIBUTOR

ITO, Technology engineering and operations

4 Volume and Speed of AI Output Beyond the Controls

Review volume and change frequency describe two different capacity demands on control. Review-capacity ratings were associated with the wider score, while the written volume pattern was less clearly differentiated by the composite. Change frequency showed a stronger adjusted engagement association than its raw association. The relevant governance question is the fit between work and control capacity, with model dependence stated explicitly.

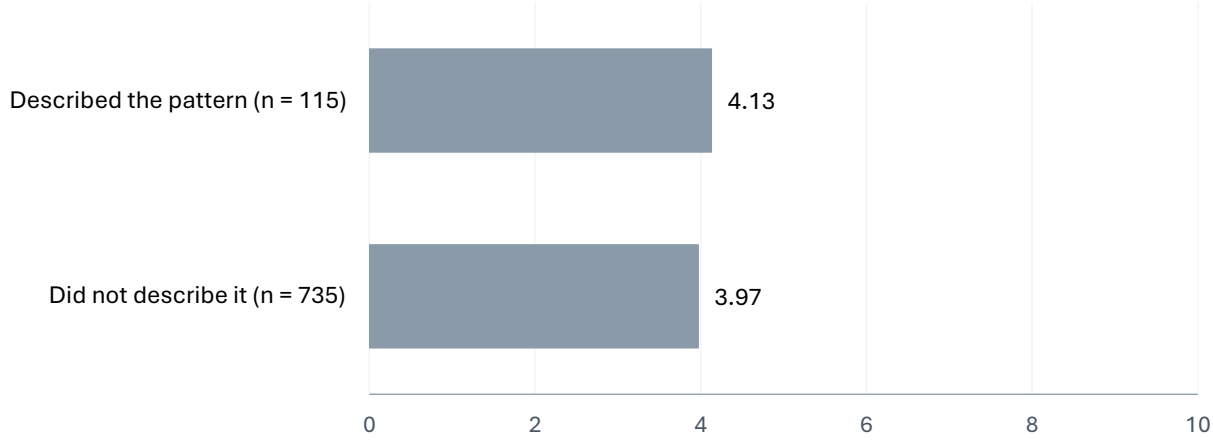
4.1 Human Review Overrun by Agent Volume

Written pattern · Answers scored negative on human review capacity, read in full · 123 answers from 115 respondents on 80 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 6.3

Human review overwhelmed by agent volume was described across a range of composite scores. The operational concern is whether review capacity can examine the work it is expected to control. The written source consists of answers classified negative on that capacity, so the pattern is not an independent survey of every respondent's review workload. A high composite does not establish that the reported capacity gap was otherwise contained.

Figure 3.94. Human Review Overrun by Agent Volume: descriptive mean score comparison.

Human review overwhelmed by agent volume was described across a range of composite scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 115 respondents who described it stood at 4.13 and the other 735 at 3.97 (cluster-robust p = 0.409). Among individual contributors alone, the 93 who described it stood at 3.93 against 3.85 (cluster-robust p = 0.667).

At the engagement level the share describing the pattern was associated with the score at r = 0.07 (p = 0.369, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = 0.00 (p = 0.954).

The pattern was described by 12.01% (34 of 283) of respondents with a score of 3 or below, 14.58% (77 of 528) above 3 to 7, and 4 of the 39 above 7. It was described by 17.14% of respondents in Build-Out, falling to 13.59% in Production and 11.18% on canceled engagements, and by 12.73% in the lowest indicator quintile against 13.12% in the highest. No role or segment family differed after Holm adjustment.

The 123 answers are the written answers scored negative on human review capacity, each read in full, and 122 of them state that review was overrun by the volume of agent output or by its own capacity. Eight of them also describe review reduced to a tick, a click-through, or a small sample, too few to stand as a pattern of their own. Thirty-six of the 123 were written on canceled engagements. On one of the five engagements without a single distressed governance rating, an individual contributor reported that the agents produced more test scripts in a morning than the team could read.

The rating of review capacity is analyzed under Co-Management and Governance Across the Other Four Categories, and Part 5 (Leading Dynamics of Manual Alteration, section 6.3) reports its association with manual alteration.

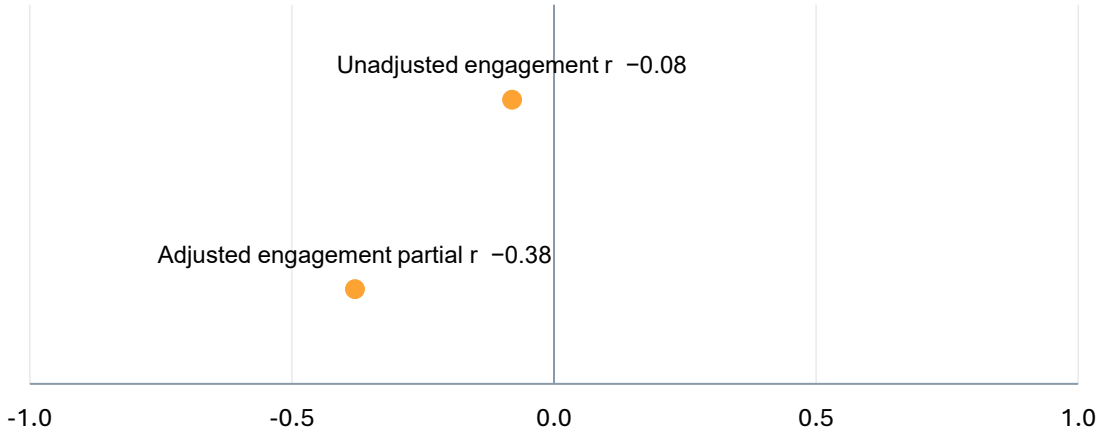
4.2 Change Frequency Beyond the Capacity of Change Control

Written pattern · Answers on the manual alteration of processes, read under the Part 5 classification of reasons · 101 answers from 93 respondents on 65 engagements

Change frequency exceeding change-control capacity had a negative adjusted association with the score in the eligible engagement comparison. The substantive concern is a process whose operating cycle cannot accommodate the changes it is meant to govern. The raw engagement correlation was small, and the respondent contrast was not clearly detected within process-answer respondents. The stronger conditional estimate depends on the specified comparison model.

Figure 3.95. Change Frequency Beyond the Capacity of Change Control: unadjusted and adjusted associations in the 133 engagements with an answer on process alteration.

Change frequency exceeding change-control capacity had a negative adjusted association with the score in the eligible engagement comparison.



Comparison	Estimate and analysis base
Unadjusted engagement r	r = −0.08 133 engagements
Adjusted engagement partial r	partial r = −0.38 133 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

Raw p = 0.366, adjusted nominal p < 0.001. The respondent comparison within 516 process-answer respondents is separate. These are conditional associations, not causal effects.

Within the 133 engagements with an answer on process manual alteration, the share describing the pattern was associated with the score at r = −0.08 (p = 0.366). After adjustment

for the mean of the other 53 indicators the association reached partial r = −0.38 (p < 0.001). Across all 150 engagements the adjusted association stood at −0.45 and held between −0.49 and −0.40 when each category's block of the 53 was omitted from the comparison mean. The share describing the pattern was associated with the 53 indicators themselves at r = −0.17.

The following distributions provide supplementary all-study context and do not replace the eligible process-answer comparisons. The pattern was described by 18.73% (53 of 283) of respondents with a score of 3 or below, 7.58% (40 of 528) above 3 to 7, and none of the 39 above 7. It was described by 14.56% of respondents in Production and 13.66% on canceled engagements, against 6.86% in Build-Out, and by 16.36% in the lowest indicator quintile against 8.75% in the highest. It differed by role after Holm adjustment, written by 92 individual contributors and one engagement manager. Among individual contributors alone, the 92 who described it stood at 3.13 against 3.98 among the other 575 (cluster-robust p < 0.001).

The 101 answers came from the answers on the manual alteration of processes, where the reading assigned each answer naming a change process, change board, queue, window, or release cycle that could not keep pace as the reason for the manual alteration, and 45 of them were written on canceled engagements. The share describing it differed after Holm adjustment by role, by management's understanding of the governing obligations, by the manual alteration methods used, and by disclosure to the financial institution. Respondents described a change process that assumed one change a month against several a day, and a queue of pending changes longer than the queue of work.

> *“The Change Process assumes one change a month and we make several a day... we cannot wait for a ruling so we rule ourselves on most exceptions”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Corporate and back-office functions

> *“Angry if im honest. The queue of pending changes is longer than the queue of actual work, and the decision arrive after the deadline it was meant to help.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Client servicing and communications

Part 5 (Manual Alteration of Financial Institution Processes, section 2.4) analyzes the same reading as a reason for process manual alteration.

5 Non-Deterministic Output and Missing Logs

Variable output and limited evidence create different assurance problems. Output that cannot be signed off and absence of a quality assurance method concern acceptance, while missing logs concern reconstruction and diagnosis. Their score relationships varied. The joint logs-and-consistency comparison describes the lowest mean score among respondents rating both poorly, with no test of causal interaction.

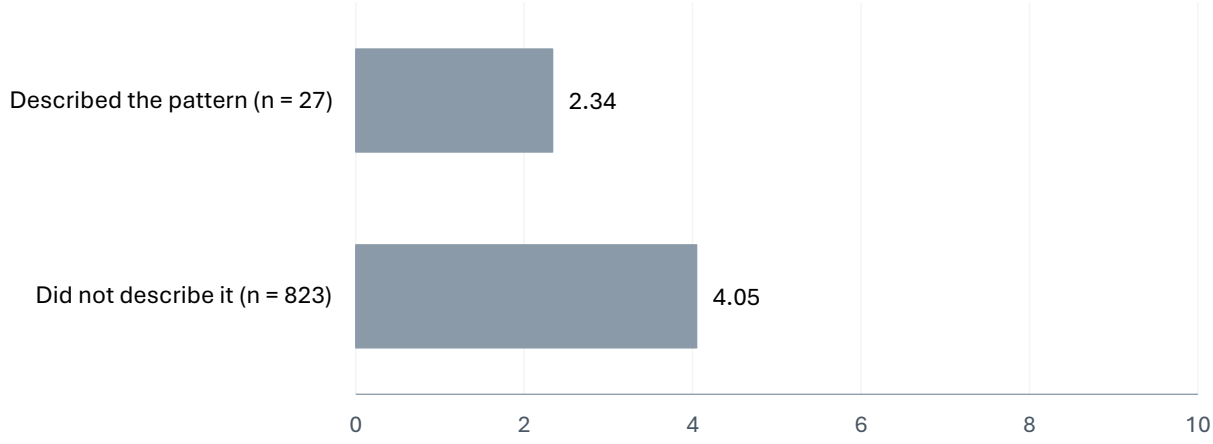
5.1 Output That Cannot Be Signed Off

Written pattern · All written answers to the open questions · 29 answers from 27 respondents on 23 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 5.4

Output that could not be signed off was described by a small group with low mean scores. This connects acceptance of agent work with the wider control environment: delivery cannot be treated as fully assured where responsible staff cannot establish a basis for sign-off. The adjusted engagement association was not clearly detected. The finding supports a specific reported assurance problem, not a causal severity ranking among written patterns.

Figure 3.96. Output That Cannot Be Signed Off: descriptive mean score comparison.

Output that could not be signed off was described by a small group with low mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 27 respondents who described it, a small base, recorded a score of 2.34, against 4.05 among the other 823 (cluster-robust $p < 0.001$). Among individual contributors alone, the 26 who described it stood at 2.22 against 3.93 (cluster-robust $p < 0.001$).

At the engagement level the share describing the pattern was associated with the score at $r = -0.32$ ($p < 0.001$, $n = 150$ engagements). After adjustment for the mean of the other 53 indicators the partial r was −0.11 ($p = 0.180$). The adjusted comparison did not detect a clear association beyond the comparison mean.

The pattern was described by 7.77% (22 of 283) of respondents with a score of 3 or below, 0.95% (5 of 528) above 3 to 7, and none of the 39 above 7. It was described by 5.28% of respondents on canceled engagements and by none in Production, and by 5.45% in the lowest indicator quintile against none in the highest. The share differed after Holm adjustment by indicator quintile, by the manual alteration methods used, and by intent to leave.

The 29 answers came from all written answers to the open questions, and 18 of them were written on canceled engagements. Part 5 counts 28 answers from 26 respondents on 22 engagements for the same pattern, among the written answers on the manual alteration methods alone, so the two counts rest on different bases. Respondents described output that gave a different figure on every run and a delivery team that chose among the runs, the best of three or the middle value, before submitting one.

> *“We get duplicates, have to guess which is real, quiet from them we take as a yes, won't sign off on output that changes each rerun, pick one and go”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Credit, underwriting and suitability

> *“the problem is we cannot sign off on something that will not reproduce, so i pick the best run out of three and that's the one we submit”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Payments, cards and disputes

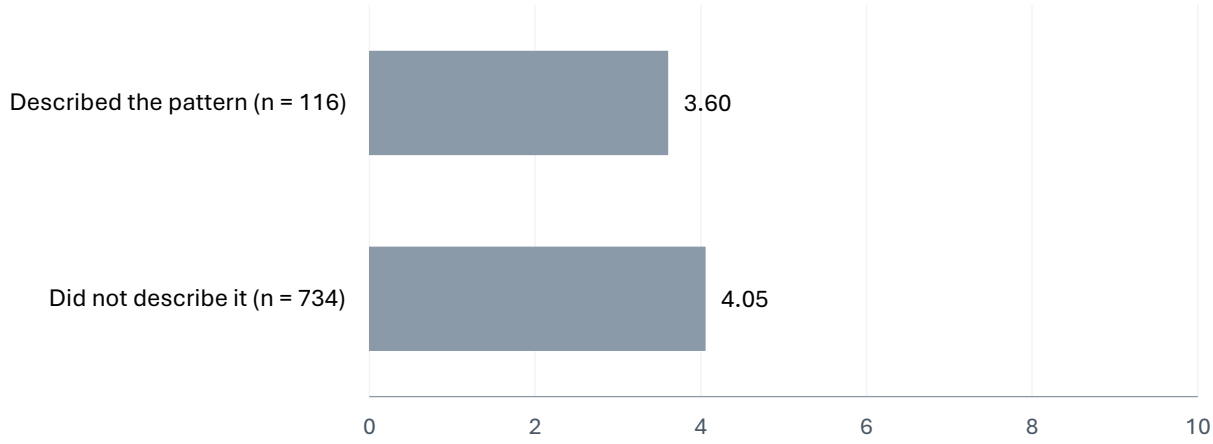
Part 5 (Leading Dynamics of Manual Alteration, section 5.4) analyzes the pattern as a factor in manual alteration.

5.2 Absence of a Quality Assurance Method for Variable Output

Written pattern · Answers scored negative on the vendor's quality assurance for variable output · 126 answers from 116 respondents on 81 engagements · Crossed with manual alteration in Part 5, Leading Dynamics of Manual Alteration, section 6.1

Absence of a quality assurance method for variable output was associated with lower respondent mean scores. Quality assurance needs a defensible basis for accepting outputs that differ across runs. The pattern comes from answers classified negative on that capability, with a less conclusive engagement-level association. It should remain distinct from the tools' testability rating and from the separate evidence provided by logs.

Figure 3.97. Absence of a Quality Assurance Method for Variable Output: descriptive mean score comparison.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 116 respondents describing the pattern had a mean score of 3.60, compared with 4.05 for the other 734 (cluster-robust p = 0.010). Among individual contributors alone, the 93 who described it stood at 3.39 against 3.94 among the other 574 (cluster-robust p = 0.005).

At the engagement level the share describing the pattern was associated with the score at r = −0.16 (p = 0.055, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.11 (p = 0.178).

The pattern was described by 18.73% (53 of 283) of respondents with a score of 3 or below, 10.80% (57 of 528) above 3 to 7, and 6 of the 39 above 7, so it reached the highest-scoring respondents as well. It was described by 16.77% of respondents on canceled engagements

against 9.71% to 13.20% across the active stages, and by 17.58% in the lowest indicator quintile against 11.25% in the highest. The share differed after Holm adjustment by intent to leave and by no other segment family.

The 126 answers are the written answers scored negative on the vendor's quality assurance for output that differs between runs, and 54 of them were written on canceled engagements. Twenty-six of them also described output that could not be signed off. Respondents described individual contributors who had become the quality assurance step, and a screenshot of every case as the only check on a tool that never gave the same answer twice.

> *“There is no QA for a tool that never give the same answer twice. I screenshot every case before I submit it. That is my QA. Nobody ask me to do it”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Financial crime and client onboarding

> *“My low marks come down to staffing, not to the leadership. We were promised two analysts with market research background and got two juniors who had never read a fund factsheet. Then QA. Nobody at my level was ever shown how to check work that comes out different every run, we werent given a method, we invented one ourselves.”*
>
> **ENGAGEMENT MANAGER**
>
> KPO, Corporate and back-office functions

The rating of the vendor's quality assurance for variable output is reported in Part 2 (Vendor Capability, section 6.2), and Part 5 (Leading Dynamics of Manual Alteration, section 6.1) reports its association with manual alteration.

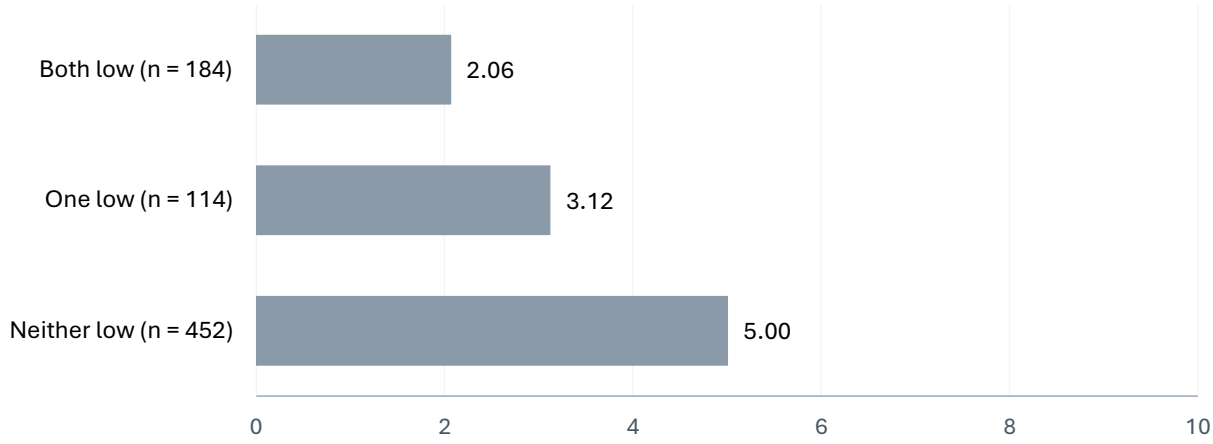
5.3 Missing Logs and Evidence of Agent Actions

Rating crossing · Logs, traces, and usage telemetry against consistency of output across runs · 750 respondents who rated both

Respondents giving low ratings to both logs and output consistency had lower mean scores than those giving a low rating to one or neither. Limited logs can make the reasons for variable results difficult to reconstruct. The comparison is descriptive and does not establish statistical interaction or causal compounding. The full composite in this comparison includes the logs item.

Figure 3.98. Logs and output consistency: mean score among 750 respondents with both numeric answers.

Respondents giving low ratings to both logs and output consistency had lower mean scores than those giving a low rating to one or neither.



Note: One hundred respondents missing at least one item are excluded. This joint description is not a test of interaction.

Among the 750 respondents with both numeric answers, 184 rated both items at 1–3 and had a mean score of 2.06. The 114 rating exactly one item at 1–3 had a mean of 3.12. The 452 rating neither at 1–3 had a mean of 5.00. The 100 respondents missing at least one answer are excluded. Across engagements, the two indicator ratings correlated at r = 0.94 (n = 150).

The ordered means describe joint reported weakness, with no test of interaction or causal compounding. The composite in this comparison includes the logs and telemetry item, so part of its relationship with the groups is built into the score. This is not independent confirmation of the logs rating. The logs item itself had an engagement-level correlation of 0.66 with the score within canceled engagements.

Written accounts concerning these limitations are examined alongside the relevant controls. Logs and telemetry drew 101 written answers and explanation at audit depth drew 163, both reported under Technical Controls and Governance Across the Engagement, and output that could not be signed off is the written form of the same failure. Part 2 (AI Technology, sections 2.3 and 8.3) reports the two ratings, and Part 5 (Leading Dynamics of Manual Alteration, sections 5.2 and 6.6) their association with manual alteration.

6 Respondent Explanations Locating Problems Upstream

Some respondents located the problem upstream of the indicator they were rating. These accounts connect a local control assessment with decisions or dependencies elsewhere in the engagement. They did not show a clearly detected adjusted score association. Their interpretive value lies in identifying the proposed location of responsibility, not in establishing the upstream account as a verified root cause.

> *Nobody gave us the real process so we reconstruct workflow by watching them work... governance forum meets monthly and we work daily so I cant wait for it"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Client servicing and communications

> *i felt there is no point giving different marks... it all come back to one thing, their compliance side will not confirm the risk rating threshold for us, therefore every case sit open. the individual people are fine, on top of that they are polite about it, but nobody will put a decision in writing"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Financial crime and client onboarding

“

We do flag it, both internally and to the client, but flagging does not stop us doing it, strictly speaking the account still needs the composites finished on time. And, management being aware does not lower how often it happens, it only means nobody is surprised the next time. It wears on the team, watching the correct process fail and having to paper over it regardless of how that looks against the rules.”

ENGAGEMENT MANAGER

KPO, Trade, settlement and reconciliation

Manual Alteration and the Operation and Visibility of Controls

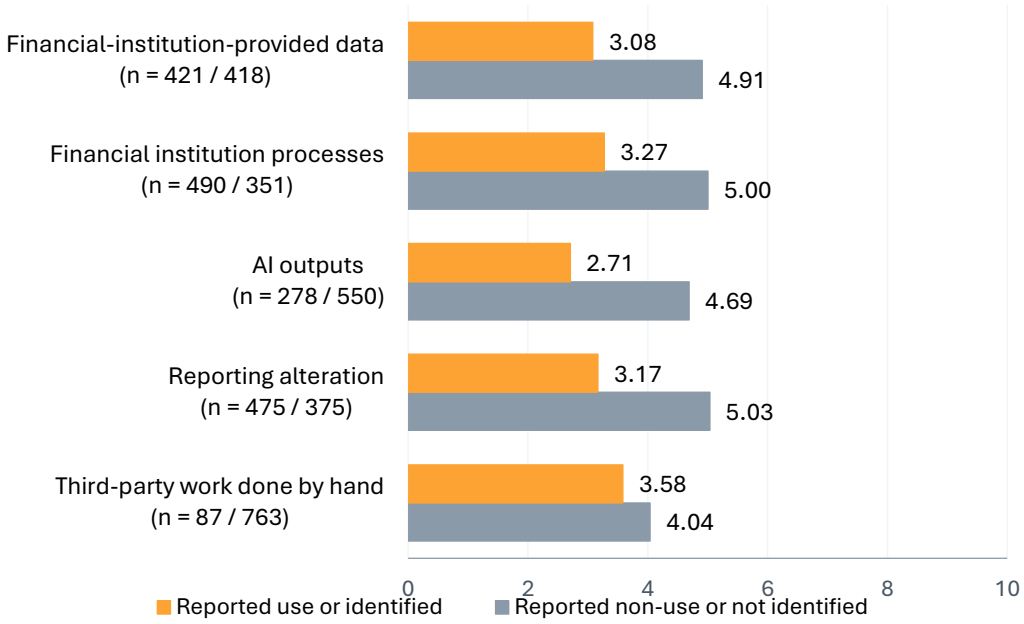
Manual work was associated with the governance, oversight, and control score, while its disclosure and recording often provided an incomplete view of how delivery operated. The five methods concern different interventions in data, processes, outputs, reporting, and third-party dependencies. Their different measurement bases and score distributions require separate interpretation. The central governance issue is whether the work remains authorized, visible, and subject to a record that oversight can examine.

The five methods are manual alteration of financial-institution-provided data, manual alteration of financial institution processes, manual alteration of AI outputs, reporting alteration, and third-party work done by hand. The first three use structured yes-or-no answers. The latter two are identified through written accounts, making a count of identified methods different from a verified inventory of all manual practice. Mean scores were 6.04 with no identified method (n = 132), 4.78 with one (210), 4.08 with two (181), 3.15 with three (147), 2.22 with four (162), and 1.85 with five (18). The group comparison was Welch $F = 271.35$ ($p < 0.001$), with $\omega^2 = 0.54$. Across engagements, the mean number identified correlated with the score at $r = -0.89$ (95% CI −0.92 to −0.85), falling to partial $r = -0.43$ after adjustment ($n = 150$, nominal $p < 0.001$).

The data, process, and AI-output methods were reported by 226 of 278, 233 of 280, and 182 of 269 answering respondents in the score band at 3 or below. Reporting alteration was identified for 238 of 283 respondents in that band. Above score 7, one respondent reported the data method, none the process method, one the AI-output method, and one described reporting alteration. Third-party work done by hand was described by 43 of 283 in the lowest band and four of 39 above 7. These observations do not certify effective controls in the upper band. Figure 3.99 compares the recorded method groups under their different ascertainment rules.

Figure 3.99. Mean score under the five methods' different recorded classifications.

Respondents reporting manual alteration had lower mean governance scores



Note: Bars show mean governance, oversight, and control scores on the 1–10 scale. In each label, the first base is reported use or identification. The second is reported non-use or absence of identification. Data, processes, and AI outputs use direct yes-or-no answers. Reporting alteration and third-party work done by hand use written identification. Absence of written identification does not confirm non-use.

> *"It happens every day, my team fixes it and moves on, we do not log how much we touch, therefore the deck looks cleaner than the underlying work actually is"*

INDIVIDUAL CONTRIBUTOR

KPO, Corporate and back-office functions

1 The Five Manual Alteration Methods and the Score

The first four manual alteration methods were substantially more concentrated in low score bands than third-party work done by hand. Structured use of the data, process, and AI-output methods was reported almost entirely by individual contributors, while reporting alteration extended to engagement managers. These findings connect operational correction and reporting visibility with different responsibilities. Numeric bands describe the distribution and do not certify effective control above a threshold.

Table 3.15. The five manual alteration methods by numeric score band and by role

Manual alteration method	The score 3 or below	Above 3 to 7	Above 7	Cramér's V (Holm p)
Financial institution data	226 of 278 (81.29%)	194 of 522 (37.16%)	1 of 39	0.46 (p < 0.001)
Financial institution processes	233 of 280 (83.21%)	257 of 522 (49.23%)	0 of 39	0.41 (p < 0.001)
AI outputs	182 of 269 (67.66%)	95 of 520 (18.27%)	1 of 39	0.51 (p < 0.001)
Reporting provided to the financial institution	238 of 283 (84.10%)	236 of 528 (44.70%)	1 of 39	0.44 (p < 0.001)
Third-party ecosystem limitations	43 of 283 (15.19%)	40 of 528 (7.58%)	4 of 39	0.12 (p = 0.003)

Table 3.15. The five manual alteration methods by numeric score band and by role / Panel B

Manual alteration method	The score, users against non-users	Users by role, individual contributors, engagement managers, portfolio directors
Financial institution data	3.08 against 4.91	418 of 659, 3 of 150, 0 of 30
Financial institution processes	3.27 against 5.00	481 of 662, 9 of 148, 0 of 31
AI outputs	2.71 against 4.69	276 of 650, 2 of 147, 0 of 31
Reporting provided to the financial institution	3.17 against 5.03	383 of 667, 86 of 150, 6 of 31
Third-party ecosystem limitations	3.58 against 4.04	69 of 667, 17 of 150, 1 of 31

Note: Data, process, and AI-output methods use respondents answering the direct yes-or-no question. Reporting alteration and third-party work done by hand are identified from written answers. Nonmention is not confirmed nonuse. The 39 respondents above score 7 are shown as counts where appropriate. Numeric bands are descriptive. Spearman rho measures the direction and strength of a rank association. Cramér's V describes categorical association, while the ordinary chi-square calculation does not adjust for respondents clustered within engagements. Holm labels identify the stated multiplicity correction, not a correction for clustering.

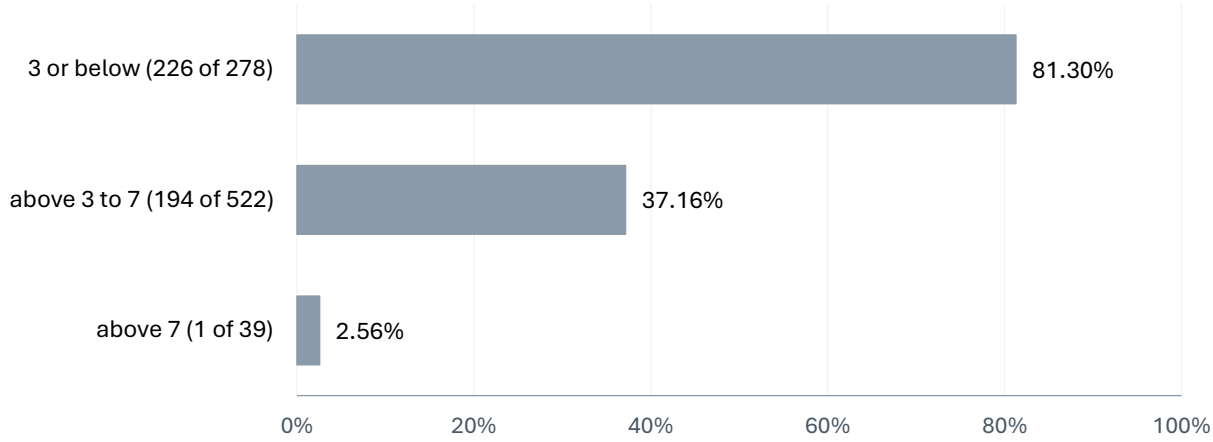
1.1 Manual Alteration of Financial-Institution-Provided Data

Manual alteration method · 421 of 839 respondents who answered · Part 5, Manual Alteration of Financial-Institution-Provided Data

Manual alteration of financial-institution-provided data was more common in the lowest score band, but substantial use also occurred in the middle band. Its governance relevance is that the delivery workforce corrects inputs that the formal input process has not made ready for agent use. The written accounts connect this work with validation, access, and response responsibilities. The distribution does not justify describing the method as confined to failed governance.

Figure 3.100. Manual Alteration of Financial-Institution-Provided Data, share of respondents using the method, by numeric score band

Manual alteration of financial-institution-provided data was more common in the lowest score band, but substantial use also occurred in the middle band.



It was used by 81.29% (226 of 278) of respondents with a score of 3 or below, 37.16% (194 of 522) above 3 to 7, and 1 of the 39 above 7 (chi-square = 178.40, Holm $p < 0.001$, Cramér's V = 0.46).

The 421 respondents who altered data recorded a score of 3.08, against 4.91 among the 418 who did not (cluster-robust $p < 0.001$). The share using the method declined across the ordered score bands at Spearman rho = −0.46 ($p < 0.001$). The gap in category governance between users and non-users held in each failure category, widest in the third-party ecosystem (−2.02) and financial institution readiness (−1.87) and narrowest in AI technology and co-management (−1.75 each).

Individual contributors accounted for 418 of the 421 users. Among individual contributors alone, 95.30% (223 of 234) with a score of 3 or below altered data, against 48.38% (194 of 401) above 3 to 7 and 1 of the 24 above 7. Three of the 150 engagement managers altered data, all three with a score of 3 or below, and none of the 30 portfolio directors who answered did.

The method's open question drew 400 written answers from 400 respondents on 127 engagements, and 213 of them were written on canceled engagements. Of the 400, 210 on 106 engagements described a governance indicator failing or a form of a control bypassed or not working. The governance indicator most often judged negative among the 210 was the financial institution's provisioning of access and identities for agents (112), ahead of its speed of response to issues raised (54) and the maturity of its data governance (49). The form most often described was no validation by the financial institution before handover (45), followed by problems raised with no response (25). The answers described data corrected before each run, from files that arrived unchecked, by staff without the access the agents needed.

> *“i get frustrated everytime, nobody update reference data in years, so i just fix it before we use, dont report it.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Corporate and back-office functions

> *“i feel stressed most days, nothing get checked on their end before it land with us, so i fix data myself before using, dont know other way.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Corporate and back-office functions

Part 5 (Manual Alteration of Financial-Institution-Provided Data) analyzes the method and its techniques in full.

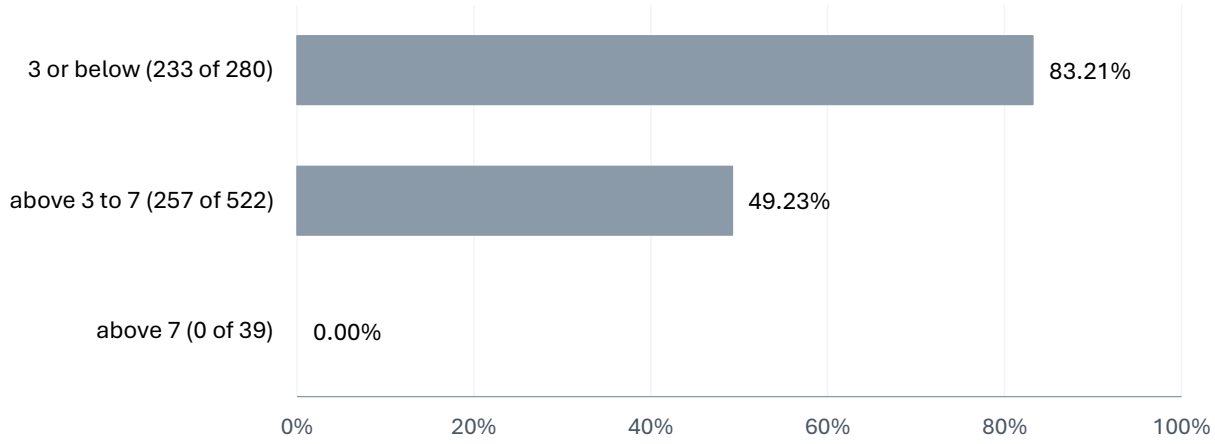
1.2 Manual Alteration of Financial Institution Processes

Manual alteration method · 490 of 841 respondents who answered · Part 5, Manual Alteration of Financial Institution Processes

Manual alteration of financial institution processes extended substantially into the middle score band. This connects reported process constraints with work performed outside the intended agent workflow. The governance question concerns whether those departures are authorized, visible, and incorporated into control arrangements. No use was observed among the 39 respondents above score 7, with no inference that process alteration cannot occur where broader ratings are favorable.

Figure 3.101. Manual Alteration of Financial Institution Processes, share of respondents using the method, by numeric score band

Manual alteration of financial institution processes extended substantially into the middle score band.



The process method was reported by 257 of 522 respondents above score 3 to 7 (49.23%). It was used by 83.21% (233 of 280) of respondents with a score of 3 or below (chi-square = 143.63, Holm $p < 0.001$, Cramér's V = 0.41).

The 490 respondents who altered processes recorded a score of 3.27, against 5.00 among the 351 who did not (cluster-robust $p < 0.001$). The share using the method declined across the ordered score bands at Spearman rho = −0.40 ($p < 0.001$). The gap between users and non-users was widest in co-management (−1.93) and financial institution readiness (−1.70) and narrowest in AI technology (−1.50).

Individual contributors accounted for 481 of the 490 users, and all 257 users above 3 to 7 were individual contributors. Among individual contributors alone, 94.51% (224 of 237) with a score of 3 or below altered processes, against 64.09% (257 of 401) above 3 to 7 and none of the 24

above 7. Nine of the 148 engagement managers who answered altered processes, all nine with a score of 3 or below, and none of the 31 portfolio directors did.

The method's open question drew 464 written answers from 464 respondents on 132 engagements, and 222 of them were written on canceled engagements. Of the 464, 277 on 120 engagements described a governance indicator failing or a form of a control bypassed or not working. The governance indicator most often judged negative among the 277 was the financial institution's speed of response to issues raised (152), followed by the effectiveness of its management of its side (61). The forms most often described were problems raised with no response (51), manual alteration not logged or only partly logged (32), and undefined handoffs between the financial institution's teams (26). The answers described decisions taken by the delivery team when no ruling came, and changed steps written up later or not at all.

> *"Escalating a break means waiting on someone who never actually rules on it, we just pick a call and run with it. And whatever that forces into the workflow gets written up later, if it gets written up at all."*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Trade, settlement and reconciliation

> *"i feel stuck waiting always, so we cannot wait for ruling, we just rule ourself and move forward, nobody stop us."*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Corporate and back-office functions

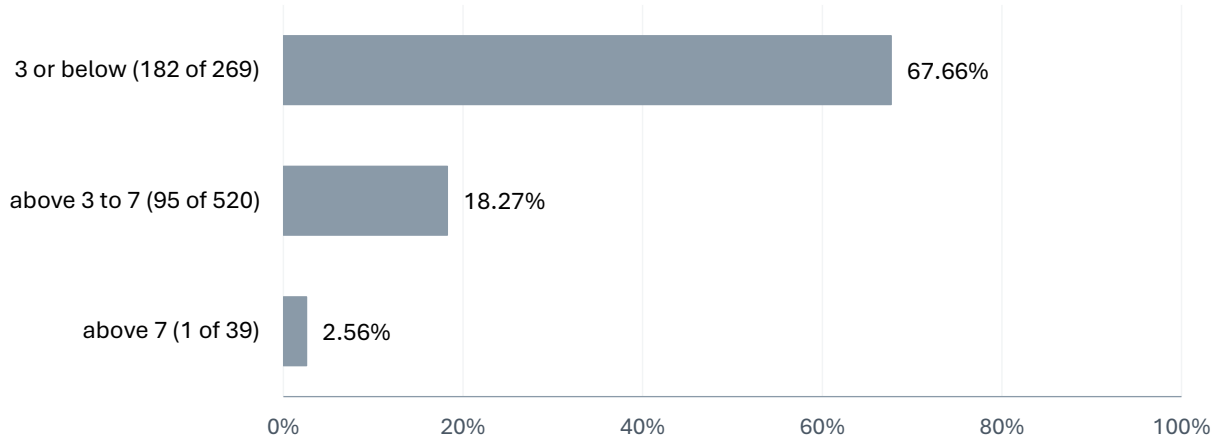
Part 5 (Manual Alteration of Financial Institution Processes) analyzes the method and its techniques in full.

1.3 Manual Alteration of AI Outputs

Manual alteration method · 278 of 828 respondents who answered · Part 5, Manual Alteration of AI Outputs

Respondents manually altering AI outputs gave lower ratings to technology-related governance controls. This connects human correction with the reported ability of the tools and surrounding processes to produce controlled work. The category gaps describe aligned assessments, rather than establish that a particular technical weakness caused each correction. The indicator remains distinct from data preparation, process changes, and alterations to the reports used for oversight.

Figure 3.102. Manual Alteration of AI Outputs, share of respondents using the method, by numeric score band



Respondents who altered AI outputs rated the AI technology governance indicators at 3.10, against 5.37 among those who did not, a gap of −2.26 and the widest of the 25 gaps between a method and a failure category. The gap in the third-party ecosystem followed at −2.19, and the narrowest gap lay in financial institution readiness (−1.55).

The method was used by 67.66% (182 of 269) of respondents with a score of 3 or below, 18.27% (95 of 520) above 3 to 7, and 1 of the 39 above 7 (chi-square = 211.55, Holm $p < 0.001$, Cramér's V = 0.51), the steepest gradient of the five methods. The 278 respondents who altered AI outputs recorded a score of 2.71, the lowest among the users of any method, against 4.69 among the 550 who did not (cluster-robust $p < 0.001$). The share using the method declined across the ordered score bands at Spearman rho = −0.50 ($p < 0.001$).

Individual contributors accounted for 276 of the 278 users. Among individual contributors alone, 79.30% (180 of 227) with a score of 3 or below altered AI outputs, against 23.81% (95 of 399) above 3 to 7 and 1 of the 24 above 7. Two of the 147 engagement managers who answered altered AI outputs, and none of the 31 portfolio directors did.

The method's open question drew 266 written answers from 266 respondents on 107 engagements, and 160 of them were written on canceled engagements. Only 72 of the 266, on 51 engagements, described a governance indicator failing or a form of a control bypassed or not working, the lowest share of the four methods asked about in an open question. Human rework absorbing agent failure appeared in 127 of the 266 answers, and the form most often described was output that cannot be signed off (19), followed by an instruction not to raise the problem (9). The answers described output that changed between runs on the same input and was corrected before anyone downstream saw it.

> *“i run it twice and get two different answers, nothing else changed. i can't trust it so i correct it before anyone sees.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Corporate and back-office functions

> *“We cannot predict what it will produce... because of that I check every output myself before it goes anywhere, and another thing this takes half my day.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> KPO, Credit, underwriting and suitability

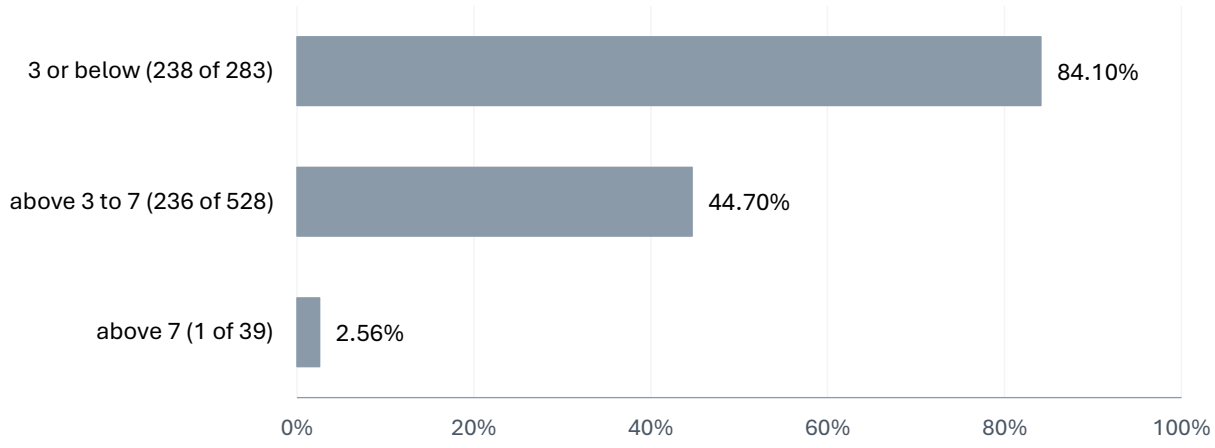
Part 5 (Manual Alteration of AI Outputs) analyzes the method and its techniques in full.

1.4 Reporting Alteration

Manual alteration method · 475 of 850 respondents with a written answer on the manual alteration of reporting · Part 5, Reporting alteration

Reporting alteration extended beyond individual contributors to engagement managers and portfolio directors. Its governance significance concerns the information through which the financial institution sees the engagement, rather than a correction to the underlying operational work. The method is identified through written accounts. Nonmention is not confirmed nonuse, and the score distribution should be read under that different ascertainment method.

Figure 3.103. Reporting alteration, share of respondents using the method, by numeric score band



Reporting alteration was described by 86 of the 150 engagement managers, including 40 of the 42 with scores at 3 or below, and six of the 31 portfolio directors. Among all respondents, it was described by 238 of 283 (84.10%) in the lowest score band and 236 of 528 (44.70%) above 3 to 7. One of the 39 above 7 described it. The ordinary band comparison was chi-square = 163.16 (Holm $p < 0.001$, Cramér's V = 0.44), with the clustering limitation stated in the methods note.

The 475 respondents who described the manual alteration of reporting recorded a score of 3.17, against 5.03 among the 375 who did not (cluster-robust $p < 0.001$). The share describing it declined across the ordered score bands at Spearman rho = −0.43 ($p < 0.001$). The gap between the two groups was widest in the third-party ecosystem (−2.05) and AI technology (−1.86) and narrowest in financial institution readiness (−1.77).

Individual contributors described the manual alteration of reporting at 57.42% (383 of 667). Among individual contributors alone, 82.43% (197 of 239) with a score of 3 or below described it, against 45.79% (185 of 404) above 3 to 7 and 1 of the 24 above 7. Among engagement managers above 3 to 7, 46 of 98 described it, and none of the 10 above 7 did.

The method's open question drew 475 written answers on 125 engagements, and 252 of them were written on canceled engagements. Of the 475, 323 on 114 engagements described a governance indicator failing or a form of a control bypassed or not working, the highest share of the four methods asked about in an open question, and 298 of the 323 described a form. Activity counted in place of outcome (109), human-corrected work counted as agent work (51), and failures removed from the base of a reported rate (49) led the forms. The reporting techniques are analyzed one by one under Oversight of Reporting Provided to the Financial Institution.

> *"The vendor's dashboard reported system uptime as the primary health metric, which stayed high even as answer accuracy collapsed, so leadership kept seeing a green status while employees kept getting wrong answers."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Corporate and back-office functions

> *"The vendor rolled uptime and ticket-closure numbers into one dashboard figure so a stalled procurement match still read as resolved, and nobody flagged the pilot until the client asked for raw supplier-match logs directly."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Corporate and back-office functions

Part 5 (Reporting alteration) analyzes the method and its techniques in full.

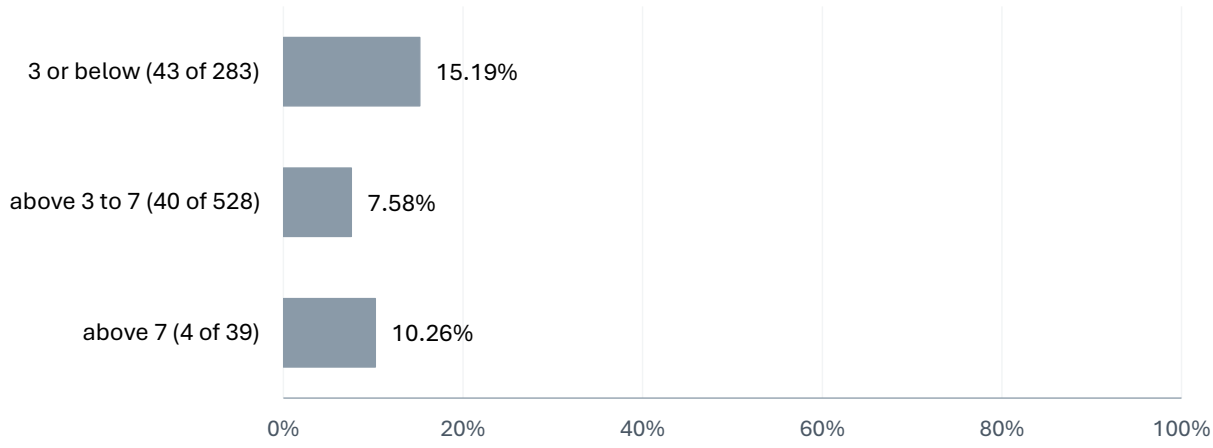
1.5 Third-Party Work Done by Hand

Manual alteration method · 87 of 850 respondents with a written answer describing manual third-party work · Part 5, Third-Party Work Done by Hand

Third-party work done by hand was less sharply concentrated in the lowest score band than the other manual alteration methods. This identifies an external-dependency constraint that can coexist with relatively favorable composite ratings. The upper band also contained isolated data, AI-output, and reporting alteration observations, so this was not the only method present there. The three-band distribution was not monotonically declining.

Figure 3.104. Third-Party Work Done by Hand, share of respondents using the method, by numeric score band

Third-party work done by hand was less sharply concentrated in the lowest score band than the other manual alteration methods.



Third-party work done by hand was described by 43 of 283 respondents at score 3 or below (15.19%), 40 of 528 above 3 to 7 (7.58%), and four of 39 above 7 (10.26%). The three-band pattern was not monotonically declining. Its categorical association was weaker than those of the other methods, with Cramér's V = 0.12, compared with 0.41–0.51 (Monte Carlo chi-square = 11.64, Holm p = 0.003). The calculation does not resolve dependence among respondents within engagements.

The 87 respondents who described it recorded a score of 3.58, against 4.04 among the other 763 (cluster-robust p = 0.042), and the trend across the ordered score bands was slight (Spearman rho = −0.11, p = 0.002). The gap between the two groups was widest in the third-party ecosystem itself (−0.76) and under 0.45 in the other four categories. At the engagement level the share describing the method was associated with the score at r = −0.14 (p = 0.097, n =

150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.05 (p = 0.544).

Engagement managers described the method at the same rate as individual contributors, 11.33% (17 of 150) against 10.34% (69 of 667). Among individual contributors alone, the shares stood at 13.39% (32 of 239) with a score of 3 or below, 8.42% (34 of 404) above 3 to 7, and 3 of the 24 above 7. The 69 individual contributors who described it stood at 3.71, against 3.88 among the other 598 (cluster-robust p = 0.474). After Holm adjustment the pattern differed only by process family and by the count of identified manual alteration methods, the second difference following in part from the definition of that count.

The 91 answers came from all written answers to the open questions, and 43 of them were written on canceled engagements. Respondents described figures pulled manually from a provider's display at each reconciliation when the custodian team did not reply, and rates pulled once a week into a workbook that every run read in place of the source.

> *But we ask custodian team and nothing come back for day, so I just pull number by hand from screen, same routine every reconciliation now”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Payments, cards and disputes

> *How we work around it is not complicated and it is not defensible either, so I may as well set it down. Nothing the agents touch on the outside is under our control, so we stopped letting them touch it. The FX and benchmark rates come down once, on a Tuesday, pulled by one of my analysts into a workbook that sits on our own drive, and every run that week reads from the workbook rather than from the source. It means the numbers are stale by design. It also means that when a provider renames a field mid-week, which happens twice in the two months we run this, the models keep producing output instead of failing, and we find out on the following Tuesday when the pull breaks, rather than at four in the morning on a deadline.I would not put this in a status report. It is a cache with a person in front of it, and the person is a qualified analyst spending Tuesday mornings on a download. But we are two months into a pilot and the alternative is a model that stops mid-run with nobody able to say why, in front of a client already halfway out the door.”*
>
> **ENGAGEMENT MANAGER**
>
> KPO, Corporate and back-office functions

Part 5 (Third-Party Work Done by Hand) analyzes the method in full.

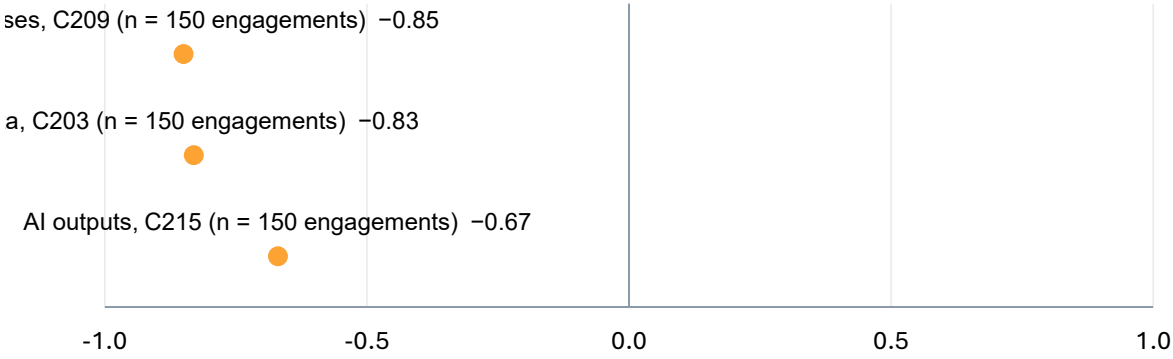
1.6 Regular Manual Alteration by the Wider Team

Rated items · Agreement that the wider team regularly altered data, processes, and AI outputs · 773, 779, and 665 respondents · Part 5, section 1.2 of each method

Ratings of regular manual alteration by the wider team were inversely associated with the score for data, processes, and AI outputs. These measures concern whether manual work is an ordinary feature of delivery rather than an isolated correction by the respondent. Their common direction connects collective practice with the reported control environment. Their separate control implications concern input preparation, departures from the intended process, and human correction of delivered outputs.

Figure 3.105. The three wider-team items: signed correlations with the engagement score.

Ratings of regular manual alteration by the wider team were inversely associated with the score for data, processes, and AI outputs.



Comparison	Estimate and analysis base
Processes, C209 (n = 150 engagements)	r = −0.85 150 engagements
Data, C203 (n = 150 engagements)	r = −0.83 150 engagements
AI outputs, C215 (n = 150 engagements)	r = −0.67 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

Negative values indicate lower scores with more regular reported manual alteration. The text reports each item and its uncertainty separately.

At the engagement level, agreement that the wider team regularly altered the financial institution's processes was associated with the score at r = −0.85 (95% CI −0.89 to −0.80).

Agreement on data stood at −0.83 (95% CI −0.87 to −0.77) and agreement on AI outputs at −0.67 (95% CI −0.75 to −0.57).

Manual alteration of financial-institution-provided data (C203). For data, agreement that the wider team regularly performed manual alteration correlated with the score at r = −0.83 across engagements (95% CI −0.87 to −0.77, n = 150). The adjusted association was partial r = −0.24 (nominal p = 0.003). At respondent level, r was −0.79 among 773 numeric answers, with the reported cluster-robust p < 0.001. The governance issue is routine input correction becoming part of collective delivery practice.

Manual alteration of financial institution processes (C209). For processes, the engagement correlation was r = −0.85 (95% CI −0.89 to −0.80, n = 150), with partial r = −0.55 after adjustment (nominal p < 0.001). The respondent correlation was −0.83 among 779 answers, with cluster-robust p < 0.001. This indicates a strong inverse association between perceived regular process intervention and the score, without establishing how much work was altered or who authorized it.

Manual alteration of AI outputs (C215). For AI outputs, the engagement correlation was r = −0.67 (95% CI −0.75 to −0.57, n = 150), with partial r = −0.30 after adjustment (nominal p < 0.001). The respondent correlation was −0.71 among 665 answers, with cluster-robust p < 0.001. The control question concerns routine human correction of the agents' delivered work and its visibility in review and reporting. No score-by-rating-band calculation is reported for these three wider-team items, so their analysis remains association based.

> *“But i do not report it, i just correct it and carry on. it happens on most files, my team all do the same, nobody raise it up officially.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Trade, settlement and reconciliation

> *“Decision rights score is one out of ten, everyone on my team feels same, nobody upstairs decides so we all just decide alone every day”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Corporate and back-office functions

> *“please understand I cannot explain why it changes, I just check twice before it goes out, this is normal now for whole team...”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Client servicing and communications

Part 5 reports the three ratings in section 1.2 of the analysis of each method.

“

A funding rule change sat with them for four days… And in that time the agent kept releasing on the old rule, so we missed the cutoff twice and i squared it by hand after. Nobody could tell me who was allowed to stop it. Moreover when they asked later what the agent did, we had nothing to show them.”

INDIVIDUAL CONTRIBUTOR

BPO, Trade, settlement and reconciliation

2 Oversight of Reporting Provided to the Financial Institution

Reporting measures connect the information received by the financial institution with the wider control environment. Reported accuracy and agreement about vendor reporting alteration retained conditional associations in the main model. Written reporting techniques generally had weaker adjusted relationships. The five-year item requires an experience-qualified base and remains a retrospective perception. Each technique identifies a different limit on the information available for oversight.

Table 3.16. The three rated reporting items and the seven reporting techniques against the score

Rated indicator or reporting technique	Base	Engagements	The score, distressed or describing against the rest (p)
Accuracy of the reports	220 of 794 distressed	150	2.28 against 4.68 (p < 0.001)
Vendor manual alteration of the reports	235 of 786 distressed	150	2.37 against 4.65 (p < 0.001)
Perceived reporting change compared with five years earlier	39 of 176 rated 8–10	150	2.65 against 4.93 (descriptive)
Activity counted in place of outcome	112 answers from 111 respondents	78	3.15 against 4.12 (p < 0.001)
Human-corrected work counted as agent work	52 answers from 52 respondents	45	3.65 against 4.02 (p = 0.113)
Failures removed from the base of a reported rate	50 answers from 50 respondents	39	3.46 against 4.03 (p = 0.017)
Measures restricted to the parseable subset	38 answers from 38 respondents	32	3.30 against 4.03 (p = 0.015)
Accuracy reported on a selected validation set	18 answers from 18 respondents	16	2.76 against 4.02 (p < 0.001)
Readiness or completion reported falsely	20 answers from 20 respondents	14	2.79 against 4.02 (p < 0.001)
Exceptions omitted from reports	23 answers from 23 respondents	20	2.67 against 4.03 (p < 0.001)

Rated indicator or reporting technique	Engagement r (p)	Adjusted partial r (p)
Accuracy of the reports	0.80 (p < 0.001)	0.24 (p = 0.003)
Vendor manual alteration of the reports	0.77 (p < 0.001)	0.19 (p = 0.019)
Perceived reporting change compared with five years earlier	−0.62 (p < 0.001)	−0.08 (p = 0.354)
Activity counted in place of outcome	−0.33 (p < 0.001)	−0.01 (p = 0.890)
Human-corrected work counted as agent work	−0.12 (p = 0.135)	0.10 (p = 0.242)
Failures removed from the base of a reported rate	−0.16 (p = 0.049)	−0.06 (p = 0.436)
Measures restricted to the parseable subset	−0.18 (p = 0.031)	0.01 (p = 0.944)

Table 3.16. The three rated reporting items and the seven reporting techniques against the score (continued)

Rated indicator or reporting technique	Base	Engagements	The score, distressed or describing against the rest (p)
Accuracy reported on a selected validation set		−0.22 (p = 0.007)	−0.02 (p = 0.797)
Readiness or completion reported falsely		−0.15 (p = 0.071)	−0.09 (p = 0.296)
Exceptions omitted from reports		−0.22 (p = 0.006)	−0.07 (p = 0.426)

Note: For rated reporting items, r is the engagement-level association with the governance, oversight, and control score, using the item's stated direction. For written techniques, r relates the engagement's share describing the pattern to the score. The usual base is 150 engagements. C054 uses only numeric answers from respondents with at least five recorded career years, and its paired engagement base is reported separately. Partial r accounts for the mean of the 53 indicators outside the score. Its p values are nominal. Written nonmention is not confirmed absence of a technique. C054 is reported in its raw direction, where higher ratings indicate more perceived reporting alteration than five years earlier. Its negative correlation is not the favorable-direction coefficient used for the other adverse item.

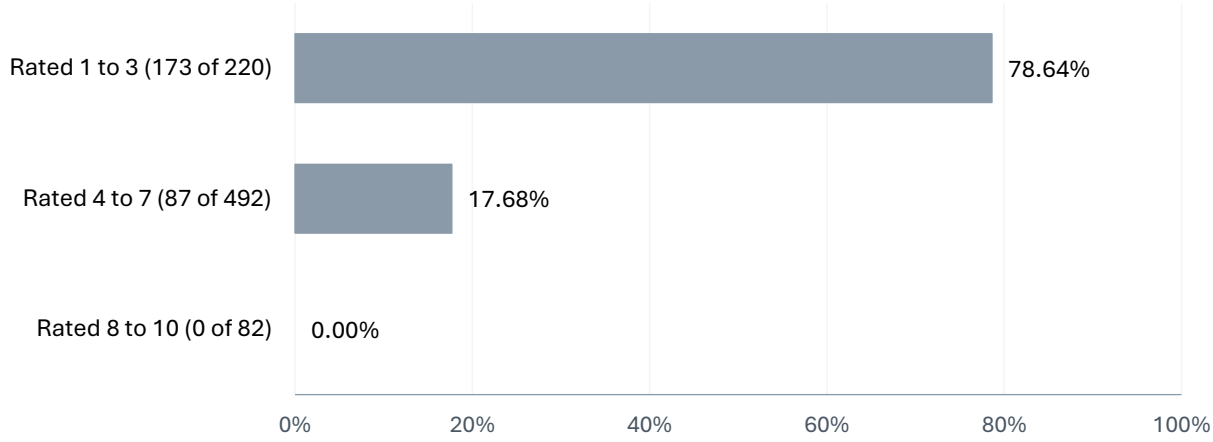
2.1 Accuracy of the Reports Given to the Financial Institution

Survey item C052 · The dashboard and performance reports given to the customer are accurate · Analyzed as a reporting measure in Part 5, Manual Alteration of Reporting Provided to the Financial Institution, section 1.1

Low reported accuracy of the information supplied to the financial institution was associated with low scores. This places the reporting channel within the governance analysis: oversight depends on a sufficiently faithful account of performance. The positive adjusted association qualifies the finding under the specified comparison model. The measure is a respondent assessment and does not independently audit the reports' arithmetic or completeness.

Figure 3.106. Accuracy of the Reports Given to the Financial Institution: share with a score of 3 or below by indicator rating.

Low reported accuracy of the information supplied to the financial institution was associated with low scores.



Note: The score uses all 47 governance indicators. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

Respondents who rated the accuracy of the reports 1 to 3 recorded a score of 2.28 (n = 220), against 4.68 among the other 574 (cluster-robust $p < 0.001$). Of the 220, 173 (78.64%) had a score of 3 or below. Among the 492 who rated accuracy 4 to 7, 87 (17.68%) did, and none of the 82 who rated it 8 to 10 did (chi-square = 300.99, $p < 0.001$, Cramér's V = 0.62).

Accuracy was associated with the score at $r = 0.80$ (95% CI 0.73 to 0.85, n = 150 engagements). After adjustment for the mean of the other 53 indicators, the adjusted estimate was partial $r = 0.24$ ($p = 0.003$), the strongest adjusted association of the three reporting items. The item sits outside the 47 governance indicators, so no cross-category measure applies.

The association was observed within the canceled engagements at r = 0.72 and across the active stages at 0.62 in Build-Out, 0.74 in Transition/Cut-Over, and 0.78 in Production. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.79 among individual contributors, 0.77 among engagement managers, and 0.66 among portfolio directors. Among individual contributors alone, 81.42% (149 of 183) of those who rated accuracy 1 to 3 had a score of 3 or below.

The item drew 468 written answers from 453 respondents on 125 engagements, and 467 of them were negative, 248 from canceled and 219 from active engagements. Of the 467, 304 (65.10%) described one of the 26 forms of a control bypassed or not working, most often activity counted in place of outcome (112), human-corrected work counted as agent work (51), and failures removed from the base of a reported rate (50). The answers described dashboards that counted throughput and left out the decisions people reversed, and savings measured against a baseline drawn from the same faulty data.

> *“The compliance dashboard tracked case throughput but not the rate of overridden agent decisions, so the volume of onboarding cases processed looked healthy while the growing number of decisions compliance staff had to quietly reverse never surfaced anywhere.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Financial crime and client onboarding

> *“The vendor's procurement dashboard reported cost savings using the agent's recommended price against an unverified prior baseline pulled from the same broken supplier master, so the savings number could not be trusted.”*
>
> **ENGAGEMENT MANAGER**
>
> CXM, Corporate and back-office functions

Part 5 (Reporting alteration, section 1.1) reports the accuracy of the reports as a reporting measure.

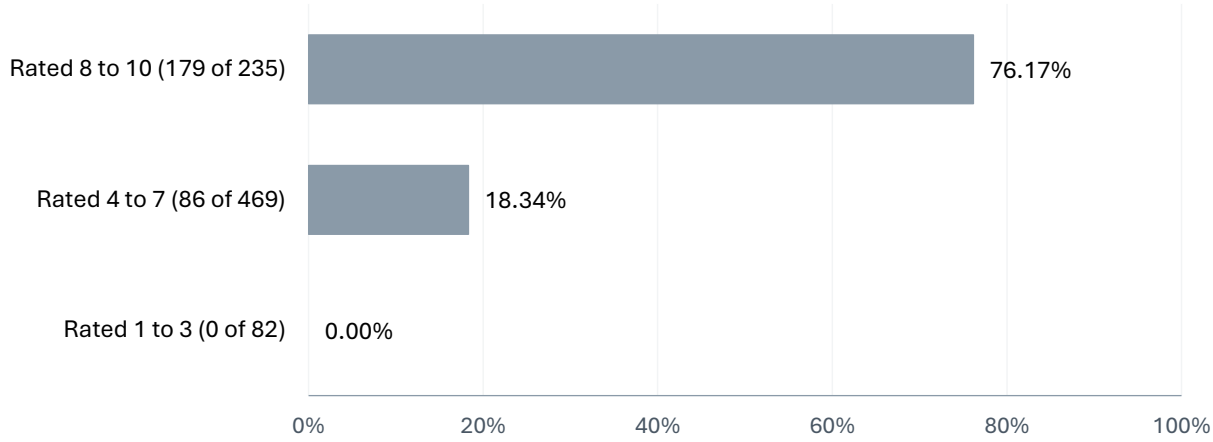
2.2 Vendor Manual Alteration of the Reports Given to the Financial Institution

Survey item C053 · The vendor manipulates the dashboard and performance reports given to the customer · Analyzed as a reporting measure in Part 5, Manual Alteration of Reporting Provided to the Financial Institution, section 1.2

Greater agreement that the vendor alters its reporting was associated with lower scores when the item's direction is interpreted consistently. This concerns intervention in the information used for oversight, distinct from correcting operational data or outputs. The reported association supports examining the integrity of that information channel. It does not establish the motive, authorization, or effect of every reporting change.

Figure 3.107. Vendor manual alteration of the Reports Given to the Financial Institution: share with a score of 3 or below by indicator rating.

Greater agreement that the vendor alters its reporting was associated with lower scores when the item's direction is interpreted consistently.



Note: The score uses all 47 governance indicators. Labels give the low-score numerator and the answering base. Zero observations in a band do not establish absence of risk.

The item was read on its favorable end, so its distressed group comprised the 235 of 786 respondents who rated the vendor's manual alteration of the reports above 7. They recorded a score of 2.37, against 4.65 among the other 551 (cluster-robust $p < 0.001$). Of the 235, 179 (76.17%) had a score of 3 or below. Among the 469 who rated the item 4 to 7, 86 (18.34%) did, and none of the 82 who rated manual alteration least likely did (chi-square = 280.87, $p < 0.001$, Cramér's V = 0.60).

The item was associated with the score at $r = 0.77$ (95% CI 0.70 to 0.83, $n = 150$ engagements). After adjustment for the mean of the other 53 indicators, the adjusted estimate was partial $r = 0.19$ ($p = 0.019$).

The association was observed within the canceled engagements at r = 0.66 and across the active stages at 0.57 in Build-Out, 0.71 in Transition/Cut-Over, and 0.78 in Production. After Holm adjustment, it was significant in 85 of the 87 segment levels tested. Within role it stood at 0.76 among individual contributors and engagement managers alike and 0.63 among portfolio directors. Among individual contributors alone, 77.89% (155 of 199) of those in the distressed group had a score of 3 or below.

The written answers on the vendor's manual alteration of the reports were largely those on their accuracy, with 460 answers judged on both items. The item drew 465 answers from 456 respondents on 125 engagements, and 464 of them were negative, 246 from canceled and 218 from active engagements. Of the 464, 304 (65.52%) described one of the 26 forms of a control bypassed or not working. One individual contributor described requisitions logged as processed with an override flag that never reached the weekly summary, and another described a benchmark for return on investment revised downward twice.

> *“Purchase requisitions the agent could not match to a valid supplier record were logged as processed with a manual override flag that never surfaced in the weekly summary the customer received.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Corporate and back-office functions

> *“The vendor reported ROI against a benchmark case that was quietly revised downward twice since the pilot began, so the same percentage improvement appeared each quarter despite the underlying baseline having shifted.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> KPO, Corporate and back-office functions

Part 5 (Reporting alteration, section 1.2) reports the vendor's manual alteration of the reports as a reporting measure.

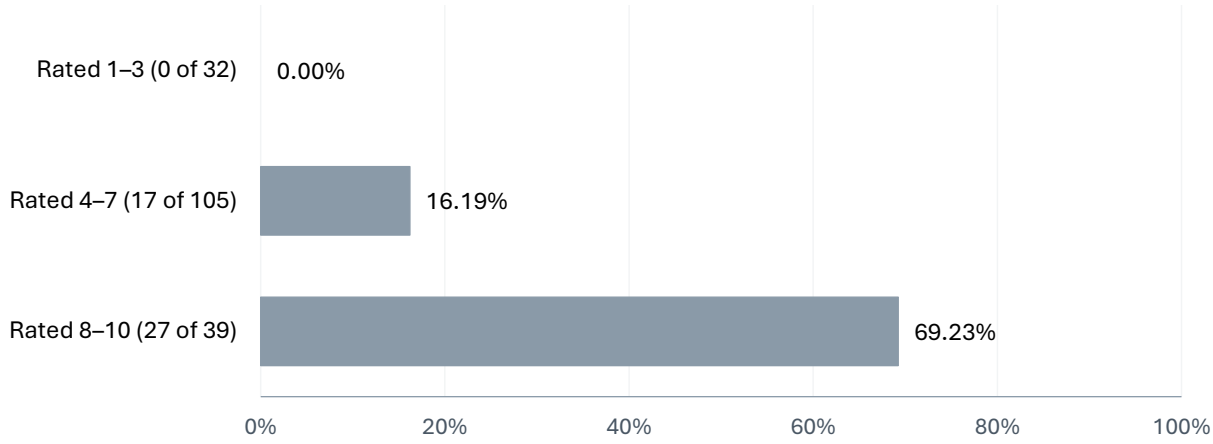
2.3 Perceived Reporting Change Compared With Five Years Earlier

Survey item C054 · Compared to 5 years ago, is manipulation of the dashboards and performance reports given to the customer more or less common now? · Analyzed as a reporting measure in Part 5, Manual Alteration of Reporting Provided to the Financial Institution, section 1.3

The five-year reporting item requires an experience-qualified interpretation. Only respondents with at least five recorded career years can support the study's intended comparison with earlier work. Their answers remain retrospective perceptions, not measured changes in reporting practice. The governance question is whether perceived reporting alteration is associated with the current control environment, with both the eligibility restriction and the paired analysis base explicit.

Figure 3.108. Perceived five-year reporting change and scores of 3 or below among 176 eligible respondents.

Experience-qualified retrospective reporting ratings differed across score groups



Note: The 567 respondents below five career years and 52 with unknown career length are excluded. Higher indicator ratings mean a perceived increase, not measured historical change.

Of the 795 numeric C054 answers, 176 met the rule of at least five recorded career years. The exclusions were 567 respondents with less than five years and 52 with unknown career length. The eligible base comprised 150 engagement managers, 24 portfolio directors, and two customer management executives, with no individual contributors. This is a substantial scope limitation. The study design allowed answers from people without the intended experience, and the eligible result describes management-level perceptions rather than the full delivery workforce.

Among eligible respondents, the 39 rating the perceived increase at 8–10 had a mean score of 2.65, compared with 4.93 among the other 137. Scores were 3 or below for 27 of those 39 (69.23%), 17 of the 105 rating the item at 4–7 (16.19%), and none of the 32 rating it at 1–3. These respondent comparisons are descriptive.

All 150 engagements had at least one eligible C054 answer. Engagement item means calculated from eligible respondents correlated with the existing full engagement score at r = −0.62 (95% CI −0.71 to −0.51, p < 0.001). Higher raw C054 ratings indicate a perceived increase in reporting alteration. After adjustment for the mean of the other 53 indicators, partial r was −0.08 (95% CI −0.23 to 0.09, nominal p = 0.354). The adjusted association was not clearly detected.

The paired engagement score and comparison mean retain their existing all-respondent definitions. Only C054's answering base is experience restricted. This distinction avoids redefining the paper's outcome for one item, but the two measures represent different respondent bases. The comparison remains retrospective perception and does not establish an actual five-year trend, its causes, or a relationship independent of management role.

> *"On the reporting engagement I ran before this one, we could never show a reviewer how a number had been arrived at. Here every step is evidenced and their audit people accepted it in the first sitting. The tool permissions still need watching."*
>
> **ENGAGEMENT MANAGER**
>
> KPO, Corporate and back-office functions

Part 5 (Reporting alteration, section 1.3) reports the change in reporting alteration as a reporting measure.

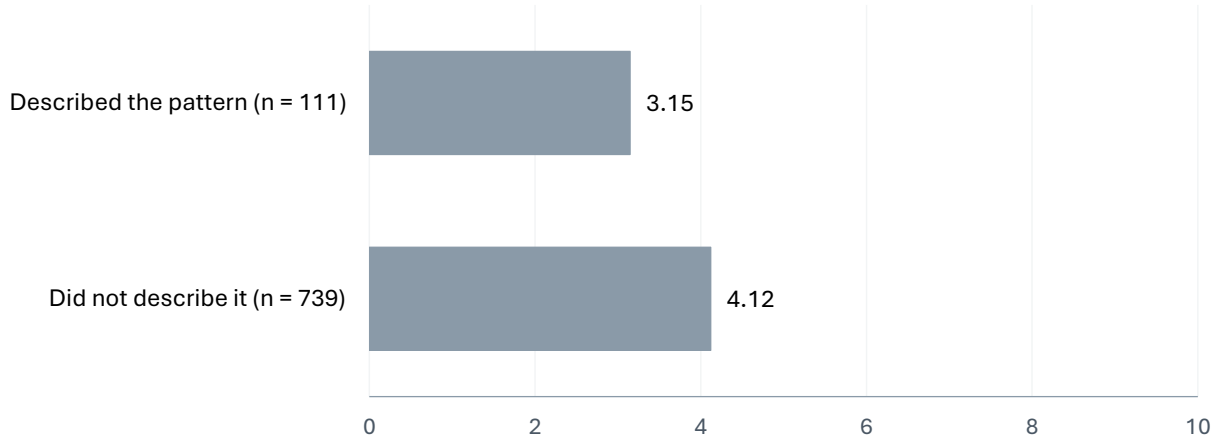
2.4 Activity Counted in Place of Outcome

Written pattern · All written answers to the open questions · 112 answers from 111 respondents on 78 engagements · Analyzed as a reporting technique in Part 5, Reporting alteration, section 2.1

Counting activity in place of outcomes was described by respondents with lower mean scores. The governance issue is whether the measure used to oversee work represents useful completion or simply recorded action. The adjusted engagement association was near zero under the main model. This limits the statistical claim while preserving the substantive difference between activity evidence and outcome evidence.

Figure 3.109. Activity Counted in Place of Outcome: descriptive mean score comparison.

Counting activity in place of outcomes was described by respondents with lower mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 111 respondents who described it recorded a score of 3.15, against 4.12 among the other 739 (cluster-robust $p < 0.001$). Among individual contributors alone, the 87 who described it stood at 3.08, against 3.98 among the other 580 (cluster-robust $p < 0.001$).

At the engagement level the share describing the technique was associated with the score at $r = -0.33$ ($p < 0.001$, $n = 150$ engagements), and after adjustment for the mean of the other 53 indicators at partial $r = -0.01$ ($p = 0.890$).

The technique was described by 20.14% (57 of 283) of respondents with a score of 3 or below, 10.23% (54 of 528) above 3 to 7, and none of the 39 above 7. It was described by 20.81% of respondents on canceled engagements, against 5.83% to 9.71% across the active stages, and

by 21.82% in the lowest indicator quintile against 5.62% in the highest. Engagement managers described it at 15.33% (23 of 150) and individual contributors at 13.04% (87 of 667), with no clearly detected difference by role after Holm adjustment. After Holm adjustment it differed by indicator quintile, management's understanding of the obligations, the financial institution's discussion of them, regulatory training, the count of identified manual alteration methods, and intent to leave.

The 112 answers came from all written answers to the open questions, and 68 of them were written on canceled engagements. Respondents described oversight counted by the transactions routed for review rather than those reviewed, and deliverables counted as shipped on schedule with no field for those returned after review.

> *"The vendor reports human-in-the-loop oversight coverage as a percentage of transactions routed for review, not the percentage actually reviewed before the transaction cleared, so a backlog of unreviewed items never dents the dashboard number."*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Payments, cards and disputes

> *"The vendor's status reports tracked deliverables shipped on schedule without a field for how many were later flagged and reworked after banker review, so the on-time metric stayed green while trust in the output was already collapsing."*
>
> **INDIVIDUAL CONTRIBUTOR**
> KPO, Corporate and back-office functions

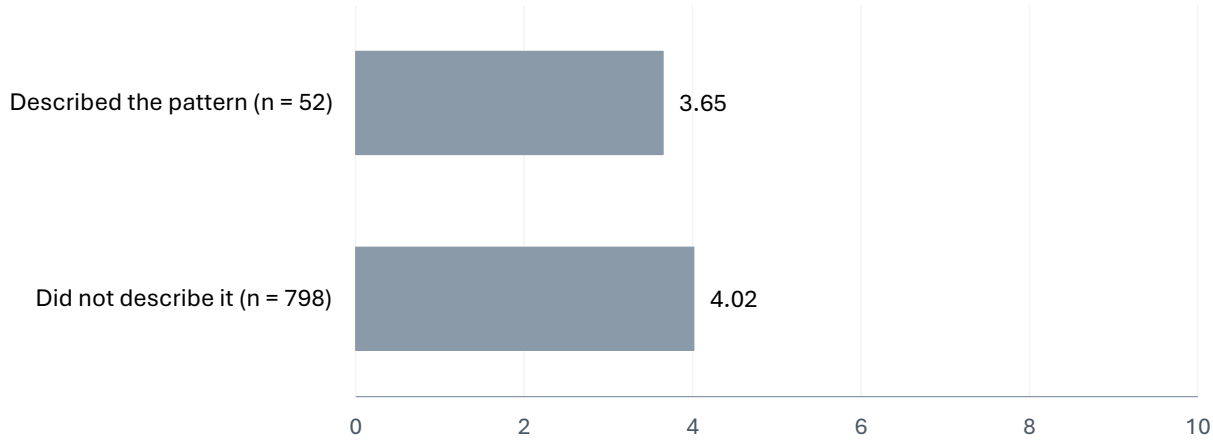
Part 5 (Reporting alteration, section 2.1) analyzes the technique under its own reading of the answers on the manual alteration of reporting.

2.5 Human-Corrected Work Counted as Agent Work

Written pattern · All written answers to the open questions · 52 answers from 52 respondents on 45 engagements · Analyzed as a reporting technique in Part 5, Reporting alteration, section 2.2

Accounts of human-corrected work being credited as agent work did not show a clearly detected score relationship. Their governance relevance lies in visibility of the human contribution to reported performance. A composite score and a description of reporting practice measure different things. The pattern remains a distinct finding even where the analysis does not establish a separate association with the broader score.

Figure 3.110. Human-Corrected Work Counted as Agent Work: descriptive mean score comparison.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 52 respondents who described it recorded a score of 3.65, against 4.02 among the other 798 (cluster-robust p = 0.113). Among individual contributors alone, the 44 who described it stood at 3.55, against 3.89 among the other 623 (cluster-robust p = 0.190).

At the engagement level the share describing the technique was associated with the score at r = −0.12 (p = 0.135, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = 0.10 (p = 0.242).

The technique was described by 7.07% (20 of 283) of respondents with a score of 3 or below, 5.87% (31 of 528) above 3 to 7, and 1 of the 39 above 7. It was described by 7.76% of respondents on canceled engagements against 4.57% to 5.60% across the active stages, and

by individual contributors at 6.60% (44 of 667) and engagement managers at 5.33% (8 of 150). It differed in no segment family after Holm adjustment.

The 52 answers came from all written answers to the open questions, and 25 of them were written on canceled engagements. Respondents described quality pass rates calculated after a reviewer had already corrected the output, and fails corrected by a person counted with the fails the agent had resolved on its own.

> *"The vendor reported QA pass rates against agent outputs that had already been silently corrected by a human reviewer before scoring, so the dashboard never reflected the raw error rate the client was actually exposed to."*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Client servicing and communications

> *"The vendor rolled settlement fails that were manually corrected by a human into the same success bucket as fails the agent resolved on its own, which overstated how much the agent was actually doing correctly."*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Trade, settlement and reconciliation

Part 5 (Reporting alteration, section 2.2) analyzes the technique under its own reading of the answers on the manual alteration of reporting.

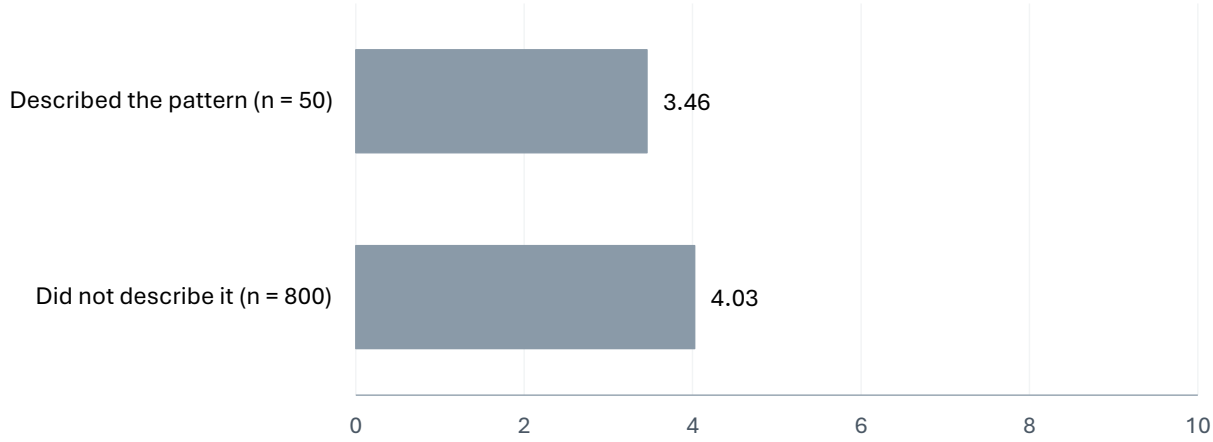
2.6 Failures Removed From the Base of a Reported Rate

Written pattern · All written answers to the open questions · 50 answers from 50 respondents on 39 engagements · Analyzed as a reporting technique in Part 5, Reporting alteration, section 2.3

Removing failures from the denominator of a reported rate was associated with lower mean scores among the respondents describing it. This affects the population to which an oversight measure refers: excluding unsuccessful work can change what a rate tells the financial institution. The overall respondent contrast, individual-contributor contrast, and adjusted engagement result differ in strength. Those differences should remain explicit rather than being compressed into one assertion of failure.

Figure 3.111. Failures Removed From the Base of a Reported Rate: descriptive mean score comparison.

Removing failures from the denominator of a reported rate was associated with lower mean scores among the respondents describing it.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 50 respondents who described it recorded a score of 3.46, against 4.03 among the other 800 (cluster-robust p = 0.017). Among individual contributors alone, the 37 who described it stood at 3.43, against 3.89 among the other 630 (cluster-robust p = 0.111).

At the engagement level the share describing the technique was associated with the score at r = −0.16 (p = 0.049, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.06 (p = 0.436).

The technique was described by 7.42% (21 of 283) of respondents with a score of 3 or below, 5.49% (29 of 528) above 3 to 7, and none of the 39 above 7. It was described by 6.86% of

respondents in Build-Out, 5.60% in Transition/Cut-Over, 3.88% in Production, and 6.21% on canceled engagements, so it was no more frequent on canceled engagements than in Build-Out. Engagement managers described it at 8.67% (13 of 150) and individual contributors at 5.55% (37 of 667), and no segment family differed after Holm adjustment.

The 50 answers came from all written answers to the open questions, and 20 of them were written on canceled engagements. Respondents described accuracy rates that left out each recommendation the agent flagged as low in confidence, and release success rates that left out any release rolled back within the first hour.

> *"Underwriting recommendations the agent flagged as low-confidence get excluded from the accuracy metric entirely, which raises the reported accuracy rate above what the full recommendation set would show."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Credit, underwriting and suitability

> *"Release-governance dashboards report deployment success against a target that silently excludes any release requiring a rollback within the first hour, so the failure rate the customer sees is well below what actually happened in production."*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Payments, cards and disputes

Part 5 (Reporting alteration, section 2.3) analyzes the technique under its own reading of the answers on the manual alteration of reporting.

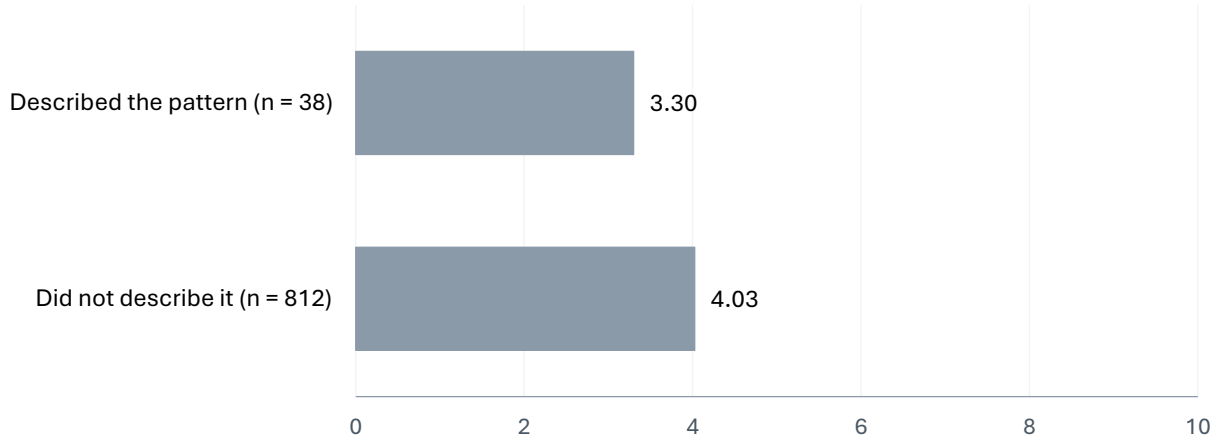
2.7 Measures Restricted to the Parseable Subset

Written pattern · All written answers to the open questions · 38 answers from 38 respondents on 32 engagements · Analyzed as a reporting technique in Part 5, Reporting alteration, section 2.4

Restricting measurement to parseable cases was described by respondents with lower mean scores. A result calculated only from parseable cases describes the work the measurement process can read. The adjusted engagement analysis did not detect a clear separate relationship. The accounts identify a selection issue without independently establishing the arithmetic or adequacy of every affected report.

Figure 3.112. Measures Restricted to the Parseable Subset: descriptive mean score comparison.

Restricting measurement to parseable cases was described by respondents with lower mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 38 respondents who described the technique recorded a score of 3.30, against 4.03 among the other 812 (cluster-robust p = 0.015). Among individual contributors alone, the 34 who described it stood at 3.27, against 3.90 among the other 633 (cluster-robust p = 0.057).

At the engagement level the share describing the technique was associated with the score at r = −0.18 (p = 0.031, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = 0.01 (p = 0.944).

The technique was described by 6.01% (17 of 283) of respondents with a score of 3 or below, 3.98% (21 of 528) above 3 to 7, and none of the 39 above 7. It was described by 5.90% of respondents on canceled engagements against 2.29% to 4.40% across the active stages, and

by 5.45% in the lowest indicator quintile against 1.25% in the highest. Individual contributors described it at 5.10% (34 of 667) and engagement managers at 2.67% (4 of 150), and no segment family differed after Holm adjustment.

The 38 answers came from all written answers to the open questions, and 19 of them were written on canceled engagements. Respondents described compliance with data residency checked once at ingestion, and access trails that logged only the grants made through the standard workflow.

> *"The vendor logs data residency compliance by checking the storage region at ingestion only, so data later copied or cached outside that region during processing never shows up as a violation on the dashboard."*
>
> **INDIVIDUAL CONTRIBUTOR**
> KPO, Corporate and back-office functions

> *"The vendor's access-control dashboard only logged provisioning events initiated through the standard workflow, so grants pushed through an emergency change process during the incident never appeared in the audit trail the bank later tried to reconstruct."*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Financial crime and client onboarding

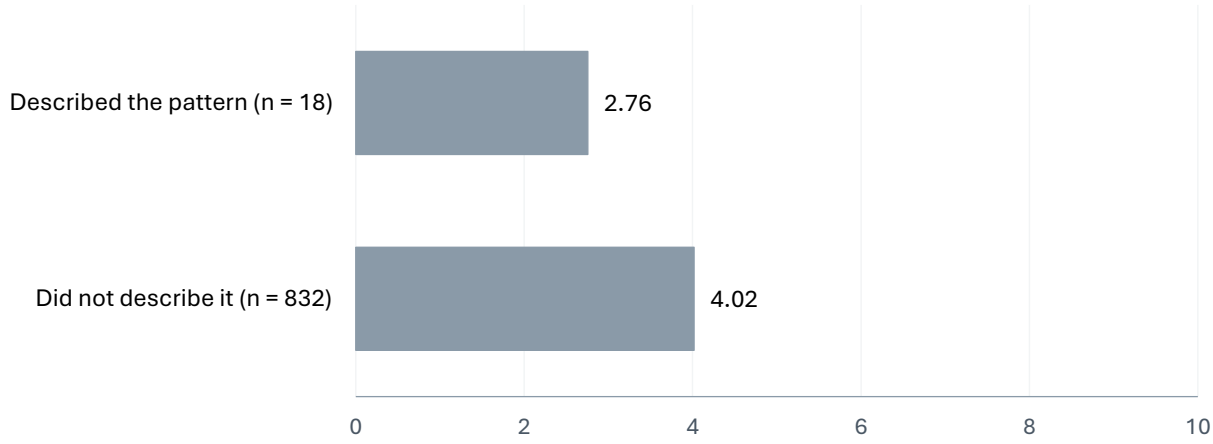
Part 5 (Reporting alteration, section 2.4) analyzes the technique under its own reading of the answers on the manual alteration of reporting.

2.8 Accuracy Reported on a Validation Set Excluding the Hardest Cases

Written pattern · All written answers to the open questions · 18 answers from 18 respondents on 16 engagements · Analyzed as a reporting technique in Part 5, Reporting alteration, section 2.5

Selected validation sets that excluded difficult cases were described by a small group with low mean scores. Their governance relevance is the relationship between evidence used to demonstrate readiness and the work the agents must perform in operation. The accounts do not independently audit the validation sets or estimate their error rates. The small written base and the inconclusive adjusted association limit broader inference.

Figure 3.113. Accuracy Reported on a Validation Set Excluding the Hardest Cases: descriptive mean score comparison.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 18 respondents who described it, a small base, recorded a score of 2.76, against 4.02 among the other 832 (cluster-robust $p < 0.001$). Among individual contributors alone, the 11 who described it stood at 2.90, against 3.88 among the other 656 (cluster-robust $p = 0.056$).

At the engagement level the share describing the technique was associated with the score at $r = -0.22$ ($p = 0.007$, $n = 150$ engagements), and after adjustment for the mean of the other 53 indicators at partial $r = -0.02$ ($p = 0.797$).

The technique was described by 11 of the 283 respondents with a score of 3 or below, 7 of the 528 above 3 to 7, and none of the 39 above 7. It was described by 3.42% of respondents on canceled engagements against 0.97% to 1.60% across the active stages, and by none of the

160 respondents in the highest indicator quintile. Engagement managers described it at 4.67% (7 of 150) and individual contributors at 1.65% (11 of 667), with no clearly detected difference by role after Holm adjustment.

The 18 answers came from all written answers to the open questions, and 11 of them were written on canceled engagements. Respondents described accuracy measured on scripted test queries in place of live tickets, and a batch rerun on a curated subset of clean accounts before each monthly steering review.

> *"The vendor's help-desk dashboard reports agent accuracy using a sample of scripted test queries rather than real employee tickets, so the reported accuracy never reflected the failure rate we saw in production."*
>
> **ENGAGEMENT MANAGER**
>
> CXM, Corporate and back-office functions

> *"Before the monthly steering review, the vendor reruns the surveillance batch on a curated subset of accounts with clean exposure histories, so the QA sample looks stable even though the production feed is still full of legacy data gaps."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Credit, underwriting and suitability

Part 5 (Reporting alteration, section 2.5) analyzes the technique under its own reading of the answers on the manual alteration of reporting.

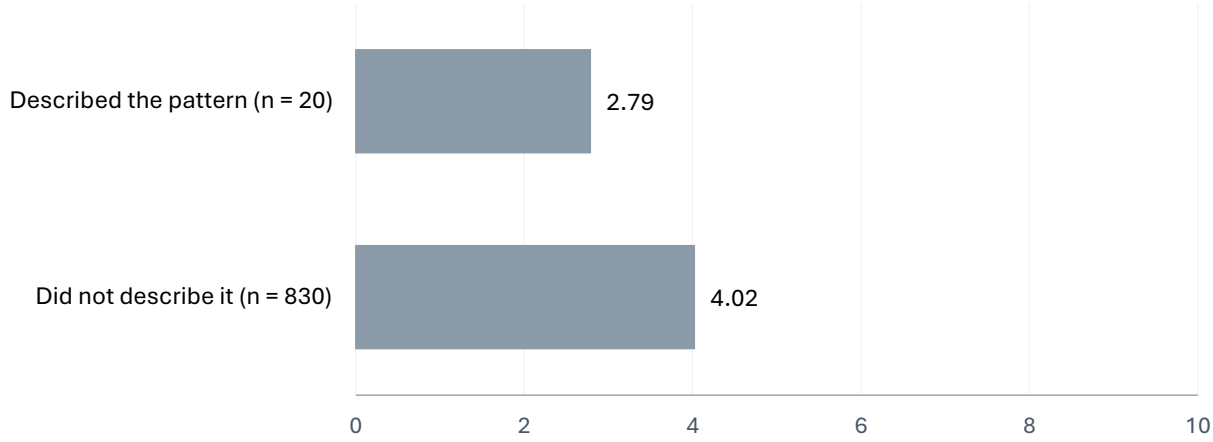
2.9 Readiness or Completion Status Reported Falsely

Written pattern · All written answers to the open questions · 20 answers from 20 respondents on 14 engagements · Analyzed as a reporting technique in Part 5, Reporting alteration, section 3.4

False readiness or completion status was described by respondents with lower mean scores, without a clearly detected corresponding engagement-level association. This is a distinct information-control concern because readiness and completion can determine whether work proceeds or is accepted. The two analytical levels describe different comparisons. Their difference should not be used to deny the respondent finding or exaggerate its engagement-wide reach.

Figure 3.114. Readiness or Completion Status Reported Falsely: descriptive mean score comparison.

False readiness or completion status was described by respondents with lower mean scores, without a clearly detected corresponding engagement-level association.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 20 respondents who described them, a small base, recorded a score of 2.79, against 4.02 among the other 830 (cluster-robust $p < 0.001$). Among individual contributors alone, the 16 who described them stood at 2.71, against 3.89 among the other 651 (cluster-robust $p = 0.004$).

At the engagement level the share describing the technique was associated with the score at $r = -0.15$ ($p = 0.071$, $n = 150$ engagements), and after adjustment for the mean of the other 53 indicators at partial $r = -0.09$ ($p = 0.296$).

The technique was described by 12 of the 283 respondents with a score of 3 or below, 8 of the 528 above 3 to 7, and none of the 39 above 7. Fourteen of the 20 respondents were on canceled engagements, where 4.35% described it against 0.40% to 2.29% across the active stages. Individual contributors described it at 2.40% (16 of 667) and engagement managers at 2.67% (4 of 150), and no segment family differed after Holm adjustment.

The 20 answers came from all written answers to the open questions, and 14 of them were written on canceled engagements. Respondents described letters marked as reviewed once a keyword filter passed, and runbooks reported as complete after the subject matter experts had flagged them as out of date.

> *"The vendor's compliance dashboard flagged drafts as 'reviewed' based on a keyword filter passing, not on a human actually reading the letter, so the one non-compliant investor letter cleared the report the same day it went out the door."*
>
> **ENGAGEMENT MANAGER**
> CXM, Client servicing and communications

> *"Leadership kept reporting the legacy runbooks as complete in the dashboard even after SMEs flagged them stale, so the readiness score stayed green while the actual documentation coverage was under half."*
>
> **ENGAGEMENT MANAGER**
> ITO, Client servicing and communications

Part 5 (Reporting alteration, section 3.4) analyzes the technique under its own reading of the answers on the manual alteration of reporting.

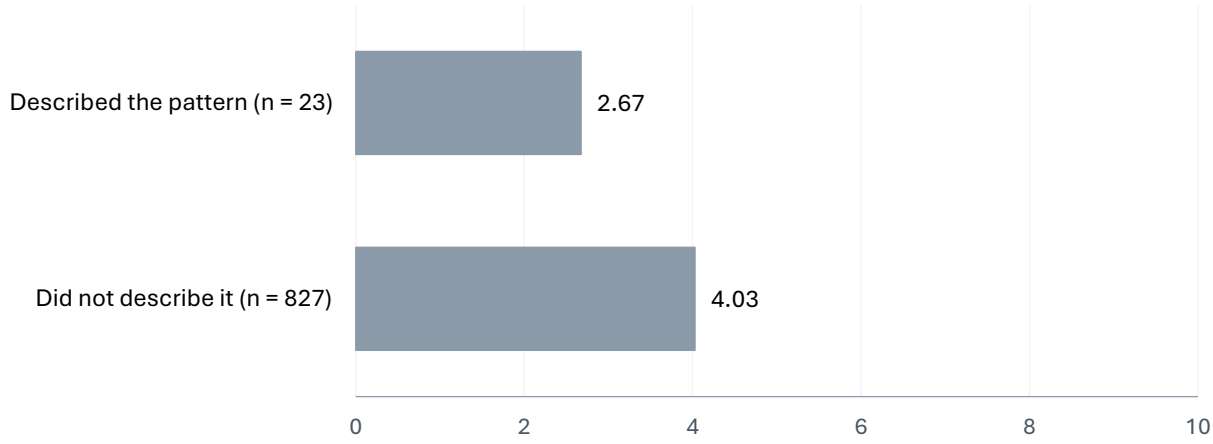
2.10 Information Withheld From the Financial Institution

Written pattern · All written answers to the open questions · 23 answers from 23 respondents on 20 engagements · Analyzed as a reporting technique in Part 5, Reporting alteration, section 4.1

Withholding recorded information from the financial institution was associated with lower mean scores among the respondents describing it. This concerns the boundary between information known within delivery and information available to the institution's oversight. It differs from a respondent simply not knowing whether disclosure occurred. The small base and the lack of a clearly detected adjusted association limit any broader inference.

Figure 3.115. Information Withheld From the Financial Institution: descriptive mean score comparison.

Withholding recorded information from the financial institution was associated with lower mean scores among the respondents describing it.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 23 respondents who described it, a small base, recorded a score of 2.67, against 4.03 among the other 827 (cluster-robust $p < 0.001$). Among individual contributors alone, the 20 who described it stood at 2.54, against 3.91 among the other 647 (cluster-robust $p < 0.001$).

At the engagement level the share describing the technique was associated with the score at $r = -0.22$ ($p = 0.006$, $n = 150$ engagements), and after adjustment for the mean of the other 53 indicators at partial $r = -0.07$ ($p = 0.426$).

The technique was described by 15 of the 283 respondents with a score of 3 or below, 8 of the 528 above 3 to 7, and none of the 39 above 7. It was described by 3.88% of respondents in Production and 3.42% on canceled engagements, against 1.14% in Build-Out and 2.40% in

Transition/Cut-Over. After Holm adjustment it differed only by the count of identified manual alteration methods.

The 23 answers came from all written answers to the open questions, and 11 of them were written on canceled engagements. Respondents described errors flagged by the financial institution routed to a separate queue that never reached the shared dashboard, and access logs kept in a console never connected to the reports a steering committee reviewed.

> *“The vendor routes any privileged-grant error flagged by the client through a separate remediation queue that never syncs back to the shared KPI dashboard, so the error rate the client sees stays lower than the actual incident count.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Technology engineering and operations

> *“Access logs for the procurement agents were kept in a separate identity-management console that was never connected to the joint reporting dashboard, so unresolved provisioning and access issues never appeared in the numbers the steering committee actually reviewed.”*
>
> **PORTFOLIO DIRECTOR**
> BPO, Corporate and back-office functions

Part 5 (Reporting alteration, section 4.1) analyzes the technique under its own reading of the answers on the manual alteration of reporting.

> *“The tool I used on the fraud desk before this was dumber but it told you when it wasnt sure. This one says done and closed and the ticket is still open underneath. Thats the part that got me. And I dont trust its own numbers on how many alerts it cleared”*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Technology engineering and operations

“

Tired, mostly… two months and we still could not get a name for who owns the vendor master file at their end. We asked for access to their procurement system four times, and received a spreadsheet extract instead. They wanted the old approval path copied exactly, agents or not. Their own people were polite and consequently I do not blame them, the gap sat above them.”

ENGAGEMENT MANAGER

CXM, Corporate and back-office functions

3 Escalation, Disclosure, and Record of Manual Alteration

Internal discussion, external disclosure, escalation, and logging describe different forms of visibility over manual work. The structured discussion items showed higher mean scores among respondents who discussed the work, while written accounts described explicit nondisclosure, unknown disclosure, and unanswered escalation. Unlogged work retained a negative adjusted association. The findings concern reported visibility, with different eligible bases and no comprehensive disclosure audit.

Table 3.17. The three items on internal discussion and the nine written patterns of disclosure, escalation, and logging against the score

Item or written pattern	Base	Engagements	The score, yes or describing against the rest (p)
Discussed data manual alteration with internal management	292 of 401 answered yes	126	3.36 against 2.40 ($p < 0.001$)
Discussed process manual alteration with internal management	341 of 459 answered yes	132	3.65 against 2.41 ($p < 0.001$)

Table 3.17. The three items on internal discussion and the nine written patterns of disclosure, escalation, and logging against the score (continued)

Item or written pattern	Base	Engagements	The score, yes or describing against the rest (p)
Discussed AI output manual alteration with internal management	150 of 253 answered yes	104	3.08 against 2.29 (p < 0.001)
Disclosure within the vendor	485 answers from 350 respondents	132	3.78 against 4.14 (p = 0.020)
Financial institution explicitly not informed	178 answers from 139 respondents	73	3.37 against 4.12 (p < 0.001)
Financial institution told	31 answers from 26 respondents	23	2.84 against 4.03 (p < 0.001)
Disclosure unknown to the respondent	30 answers from 25 respondents	20	3.18 against 4.02 (p = 0.003)
Problems raised with no response	326 answers from 255 respondents	119	3.17 against 4.35 (p < 0.001)
Raising problems seen as futile	149 answers from 122 respondents	81	2.69 against 4.21 (p < 0.001)
Instructed not to raise the problem	20 answers from 15 respondents	12	2.46 against 4.02 (p < 0.001)
Manual alteration disclosed to no party	59 answers from 55 respondents	39	2.62 against 4.09 (p < 0.001)
Manual alteration not logged or only partly logged	143 answers from 118 respondents	72	3.14 against 4.13 (p < 0.001)

Item or written pattern	Engagement r (p)	Adjusted partial r (p)
Discussed data manual alteration with internal management	0.20 (p = 0.028)	0.10 (p = 0.256)
Discussed process manual alteration with internal management	0.30 (p < 0.001)	0.08 (p = 0.339)
Discussed AI output manual alteration with internal management	0.20 (p = 0.037)	0.12 (p = 0.235)
Disclosure within the vendor	−0.19 (p = 0.023)	−0.11 (p = 0.191)
Financial institution explicitly not informed	−0.23 (p = 0.005)	0.00 (p = 0.955)
Financial institution told	−0.30 (p < 0.001)	−0.06 (p = 0.482)

Table 3.17. The three items on internal discussion and the nine written patterns of disclosure, escalation, and logging against the score (continued)

Item or written pattern	Base	Engagements	The score, yes or describing against the rest (p)
Disclosure unknown to the respondent		−0.20 (p = 0.015)	−0.03 (p = 0.728)
Problems raised with no response		−0.57 (p < 0.001)	−0.05 (p = 0.520)
Raising problems seen as futile		−0.49 (p < 0.001)	−0.22 (p = 0.008)
Instructed not to raise the problem		−0.24 (p = 0.003)	−0.04 (p = 0.588)
Manual alteration disclosed to no party		−0.30 (p < 0.001)	−0.10 (p = 0.235)
Manual alteration not logged or only partly logged		−0.30 (p < 0.001)	−0.26 (p = 0.001)

Note: The three structured discussion questions were asked only of users of the respective manual method. Respondent and engagement bases are restricted to those with an answer to that question. The nine written patterns use the stated written-answer source and all 150 engagements. Disclosure within the vendor, disclosure to the financial institution, explicit nondisclosure, and unknown disclosure are different observations. Nonmention is not a confirmed negative response. Adjusted p values are nominal unless stated otherwise.

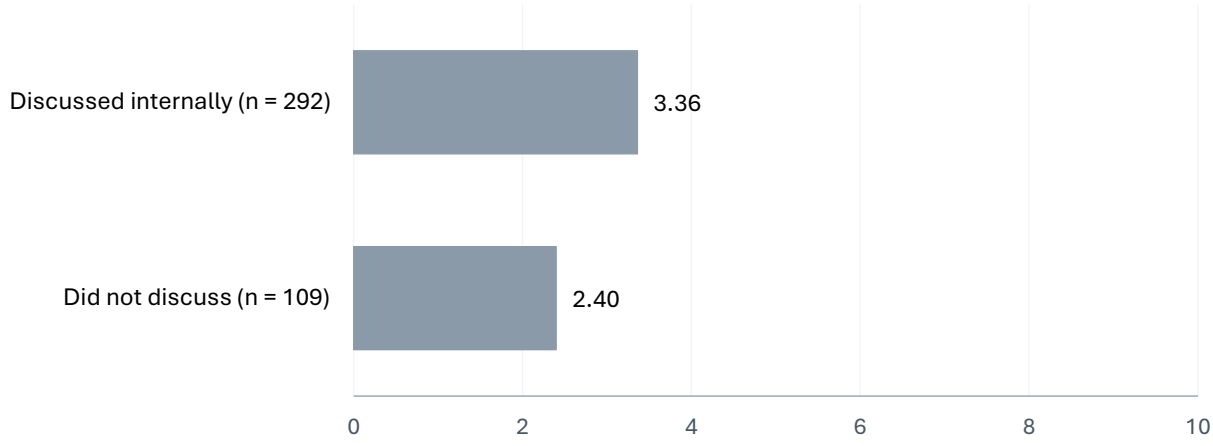
3.1 Internal Discussion of Manual Alteration of Data

Survey item C200 · I discussed the data alteration with internal management · Analyzed in Part 5, Disclosure of Manual Alteration, section 1.1

Among respondents who manually altered data, those reporting internal discussion had higher mean scores than those who did not. This connects management visibility of manual work with the reported control environment. Internal discussion is not the same as disclosure to the financial institution or formal authorization. The engagement-level adjusted result was less conclusive, and the analysis applies to the eligible method users who answered.

Figure 3.116. Internal Discussion of Manual Alteration of Data: mean score among eligible method users with an answer.

Among respondents who manually altered data, those reporting internal discussion had higher mean scores than those who did not.



Note: The engagement-level adjusted relationship is reported separately. Discussion is not a tested cause of the score difference.

The item was put only to respondents who had altered data, and 292 of the 401 who answered (72.82%) had discussed the manual alteration with internal management. They recorded a score of 3.36, against 2.40 among the 109 who had not (cluster-robust $p < 0.001$).

At the engagement level the share raising the manual alteration was associated with the score at $r = 0.20$ (95% CI 0.02 to 0.36, $p = 0.028$, $n = 126$ engagements with an answer), and after adjustment for the mean of the other 53 indicators at partial $r = 0.10$ ($p = 0.256$).

Discussion with management rose with the engagement's standing, from 67.77% (82 of 121) in the lowest indicator quintile to 93.94% (31 of 33) in the highest. It stood at 65.71% (138 of 210) on canceled engagements, against 76.47% to 84.06% across the active stages. Individual contributors gave 398 of the 401 answers, and 289 of them (72.61%) had raised the manual

alteration. After Holm adjustment the share differed by indicator quintile, engagement status, the respondent's understanding of the obligations, the financial institution's discussion of them, regulatory training, the count of identified manual alteration methods, and intent to leave.

The disclosure question for data drew 417 written answers on 127 engagements, and 221 of them were written on canceled engagements. Of the 417, 250 described a governance indicator failing or a form of a control bypassed or not working, led by the financial institution explicitly not informed (67) and problems raised with no response (59). Respondents described raising the manual alteration with a team lead and receiving an instruction to manage it, and discussing it at each retrospective while the source went unfixed.

> *“i tell my TL, kindly he say just manage it... we do the same thing every week, i keep no record of the informations i change.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Corporate and back-office functions

> *“I told my manager, not the client. Given that we just merge duplicates and move on - we discuss about it every retro but in addition nobody upstream ever fixes the actual source.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Corporate and back-office functions

Part 5 (Disclosure of Manual Alteration, section 1.1) reports the discussion of manual alteration of data with internal management.

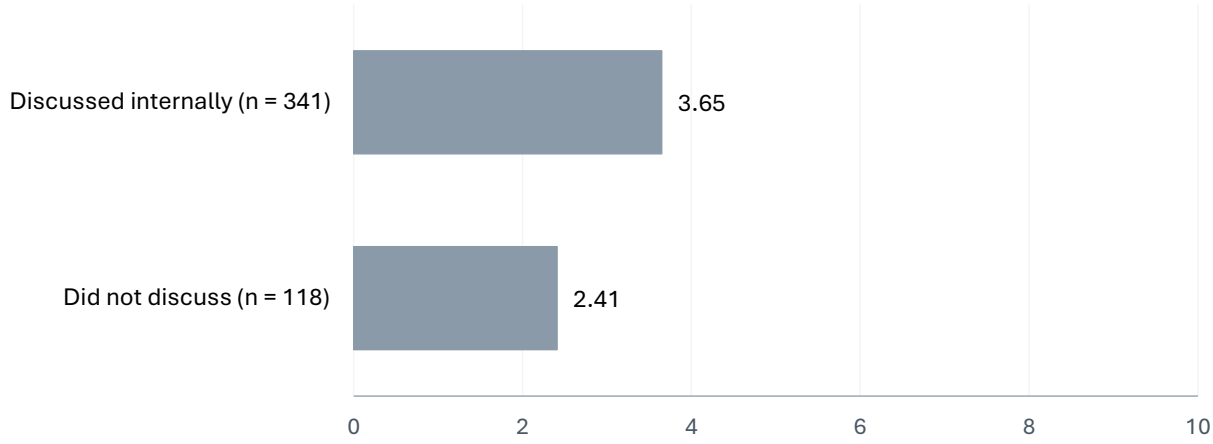
3.2 Internal Discussion of Manual Alteration of Processes

Survey item C206 · I discussed the process alteration with internal management · Analyzed in Part 5, Disclosure of Manual Alteration, section 1.2

Among respondents who manually altered processes, internal discussion was associated with higher mean scores. The governance issue is whether departures from the expected workflow become visible within the vendor's management. Discussion can be present while external disclosure or a formal record remains absent. The comparison is restricted to method users, and the adjusted engagement association was not clearly detected.

Figure 3.117. Internal Discussion of Manual Alteration of Processes: mean score among eligible method users with an answer.

Among respondents who manually altered processes, internal discussion was associated with higher mean scores.



Note: The engagement-level adjusted relationship is reported separately. Discussion is not a tested cause of the score difference.

The item was put only to respondents who had altered processes, and 341 of the 459 who answered (74.29%) had discussed the manual alteration with internal management. They recorded a score of 3.65, against 2.41 among the 118 who had not (cluster-robust $p < 0.001$).

At the engagement level the share raising the manual alteration was associated with the score at $r = 0.30$ (95% CI 0.14 to 0.45, $p < 0.001$, $n = 132$ engagements with an answer), and after adjustment for the mean of the other 53 indicators at partial $r = 0.08$ ($p = 0.339$).

Discussion with management rose from 67.44% (87 of 129) in the lowest indicator quintile to 95.35% (41 of 43) in the highest. It stood at 64.19% (138 of 215) on canceled engagements, against 81.94% to 84.25% across the active stages. Individual contributors gave 450 of the 459 answers, and all nine engagement managers who answered had raised the manual alteration.

After Holm adjustment the share differed by indicator quintile, the respondent's understanding of the obligations, the financial institution's discussion of them, regulatory training, the count of identified manual alteration methods, and intent to leave, and it did not differ by engagement status.

The disclosure question for processes drew 504 written answers on 133 engagements, and 248 of them were written on canceled engagements. Of the 504, 283 described a governance indicator failing or a form of a control bypassed or not working, led by problems raised with no response (73), the financial institution explicitly not informed (71), and manual alteration not logged or only partly logged (52). Respondents described a manager who heard of the manual alteration and did nothing, and a ticket that took days against a fix that took an hour.

> *"But this is just the same pattern every week, i tell my manager and he shrug, kindly nobody ever push back to client about undefined process"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Corporate and back-office functions

> *"This is basically every week now, my manager knows about it. Ticket takes days to clear, fix itself takes an hour, so waiting on the ticket process makes no sense when the job needs doing"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

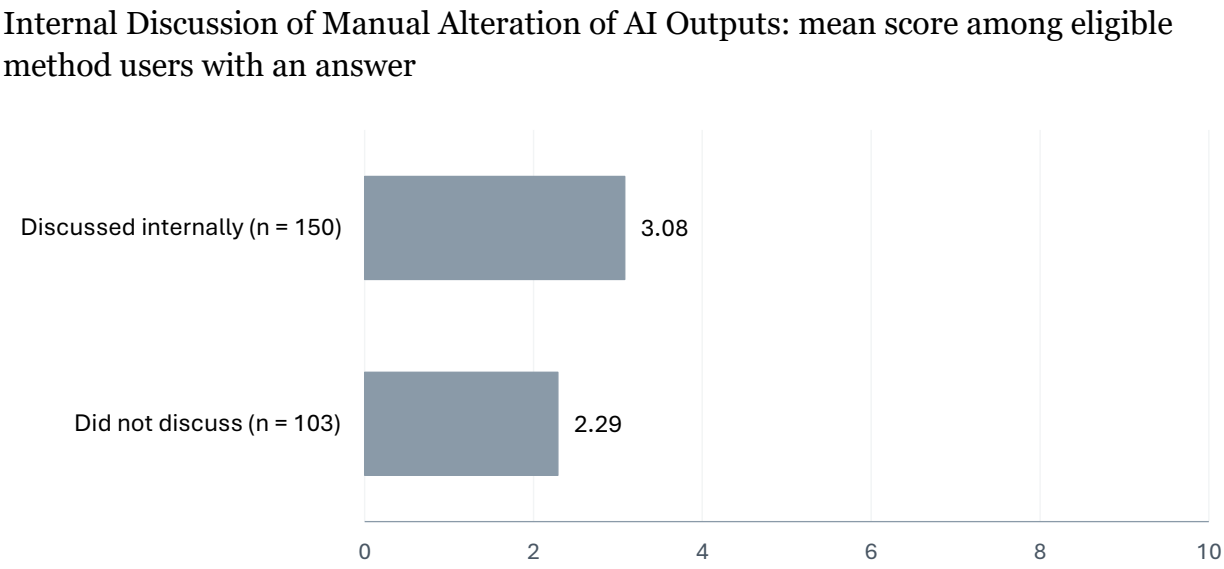
Part 5 (Disclosure of Manual Alteration, section 1.2) reports the discussion of manual alteration of processes with internal management.

3.3 Internal Discussion of Manual Alteration of AI Outputs

Survey item C212 · I discussed the AI output alteration with internal management · Analyzed in Part 5, Disclosure of Manual Alteration, section 1.3

Internal discussion of manual AI-output alteration was less commonly reported than discussion of data or process alteration. Those reporting discussion nevertheless had higher mean scores. This connects visibility of human output correction with the control environment, while leaving its authorization and external disclosure as separate questions. The findings concern the eligible users who answered and do not identify why others did not discuss the work.

Figure 3.118. Internal Discussion of Manual Alteration of AI Outputs: mean score among eligible method users with an answer.



Note: The engagement-level adjusted relationship is reported separately. Discussion is not a tested cause of the score difference.

The item was put only to respondents who had altered AI outputs, and 150 of the 253 who answered (59.29%) had discussed the manual alteration with internal management, against 72.82% for data and 74.29% for processes. They recorded a score of 3.08, against 2.29 among the 103 who had not (cluster-robust $p < 0.001$).

At the engagement level the share raising the manual alteration was associated with the score at $r = 0.20$ (95% CI 0.01 to 0.38, $p = 0.037$, $n = 104$ engagements with an answer), and after adjustment for the mean of the other 53 indicators at partial $r = 0.12$ ($p = 0.235$).

Discussion with management stood at 51.72% (45 of 87) in the lowest indicator quintile and 12 of 13 in the highest, without a steady rise between them. Across engagement status it ranged

from 54.25% (83 of 153) on canceled engagements to 82.50% (33 of 40) in Transition/Cut-Over, a difference that did not survive Holm adjustment. After Holm adjustment the share differed only by the financial institution's discussion of the obligations, regulatory training, and the count of identified manual alteration methods.

The disclosure question for AI outputs drew 274 written answers on 109 engagements, and 166 of them were written on canceled engagements. Of the 274, 115 described a governance indicator failing or a form of a control bypassed or not working, led by the financial institution explicitly not informed (29), problems raised with no response (19), and manual alteration not logged or only partly logged (19). Respondents described telling a lead each time until the instruction to note it and move on ended the raising, and a quality step that existed on paper only.

> *“i tell my lead every time because it happen a lot, kindly they just say note it and move on, it dont change anything so i stop raising the same issue.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> KPO, Corporate and back-office functions

> *“nobody reviews what i change most weeks, thus it just goes through, the qa step exists on paper only”*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Trade, settlement and reconciliation

Part 5 (Disclosure of Manual Alteration, section 1.3) reports the discussion of manual alteration of AI outputs with internal management.

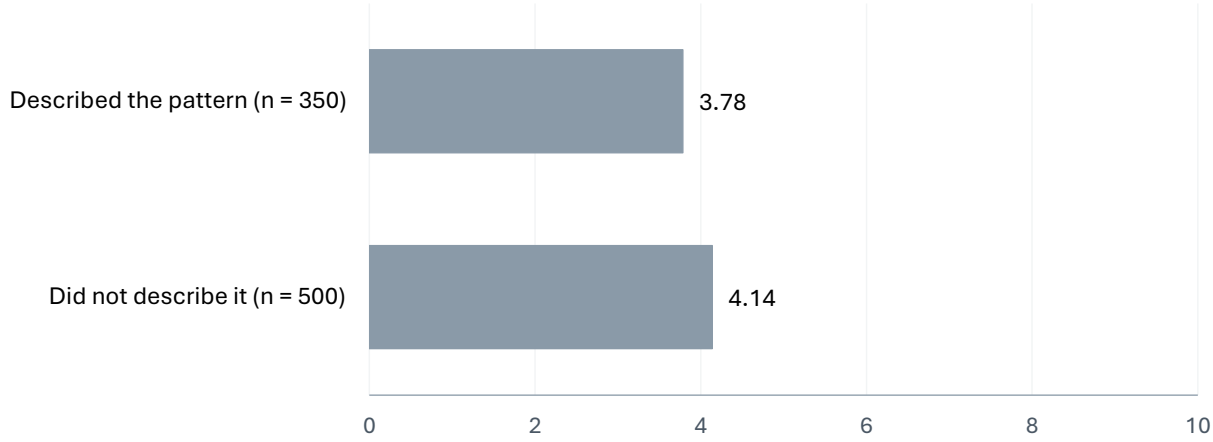
3.4 Disclosure Within the Vendor

Written pattern · All written answers to the open questions · 485 answers from 350 respondents on 132 engagements · Analyzed in Part 5, Disclosure of Manual Alteration, section 1.4

Disclosure within the vendor was associated with a lower overall respondent mean score, but the contrast was not clearly detected among individual contributors alone. This distinguishes a written disclosure pattern from the structured discussion items asked of method users. The governance interpretation concerns where manual work becomes visible and who reports it. Role composition may help explain the overall contrast, without establishing a complete explanation.

Figure 3.119. Disclosure Within the Vendor: descriptive mean score comparison.

Disclosure within the vendor was associated with a lower overall respondent mean score, but the contrast was not clearly detected among individual contributors alone.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 350 respondents who described it recorded a score of 3.78, against 4.14 among the other 500 (cluster-robust p = 0.020). Among individual contributors alone, the observed difference was small, at 3.83 among the 338 who described it against 3.90 among the other 329 (cluster-robust p = 0.706). Role composition may contribute to the overall contrast, without providing a complete explanation.

At the engagement level the share describing the pattern was associated with the score at r = −0.19 (p = 0.023, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.11 (p = 0.191).

Individual contributors described disclosure within the vendor at 50.67% (338 of 667), engagement managers at 8.00% (12 of 150), and no portfolio director did, a difference by role that held after Holm adjustment. The pattern was described at 36.89% to 42.55% across engagement status, by 41.70% (118 of 283) of respondents with a score of 3 or below and 43.94% (232 of 528) above 3 to 7, and by none of the 39 above 7. The contrast was not restated among the respondents who altered the work, so it compares all respondents.

The 485 answers came from all written answers to the open questions, and 211 of them were written on canceled engagements. Respondents described each manual alteration raised with a lead, the lead's knowledge serving as its only log, and processes rebuilt around the workaround from one sprint to the next.

> *i tell my lead each time, therefore it's logged with them at least. We chase client for weeks about the same codes and still get nothing back”*
>
> **INDIVIDUAL CONTRIBUTOR**
> ITO, Technology engineering and operations

> *i raise it with my lead only, not client side. We get informations that is already stale, so process gets built around workaround, and it stays that way every sprint.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> KPO, Corporate and back-office functions

Part 5 (Disclosure of Manual Alteration, section 1.4) analyzes manual alteration reported to vendor-side management within the answers on the three counted methods.

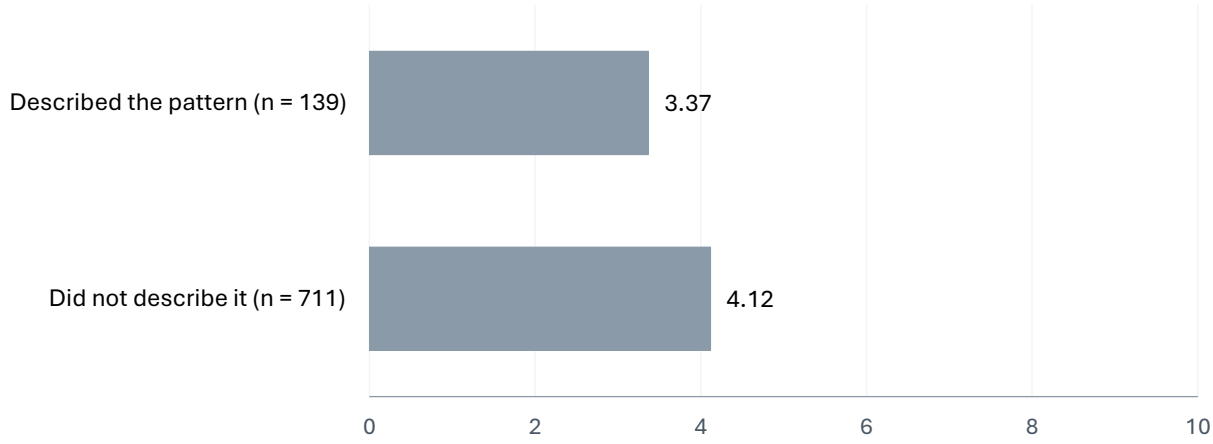
3.5 Financial Institution Explicitly Not Told

Written pattern · All written answers to the open questions · 178 answers from 139 respondents on 73 engagements · Analyzed in Part 5, Disclosure of Manual Alteration, section 2.5

Explicit accounts that the financial institution was not told were associated with lower mean scores. Reported nondisclosure identifies a possible gap between operational practice and the information available for financial institution oversight. This identifies reported nondisclosure rather than a respondent's uncertainty, while remaining short of a complete disclosure audit. The 178 written answers represent 139 respondents on 73 engagements.

Figure 3.120. Financial Institution Explicitly Not Told: descriptive mean score comparison.

Explicit accounts that the financial institution was not told were associated with lower mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 139 respondents who described the decision recorded a score of 3.37, against 4.12 among the other 711 (cluster-robust $p < 0.001$). Among individual contributors alone they stood at 3.37, against 4.00 among the other 528 (cluster-robust $p = 0.002$).

At the engagement level the share describing the pattern was associated with the score at $r = -0.23$ ($p = 0.005$, $n = 150$ engagements), and after adjustment for the mean of the other 53 indicators at partial $r = 0.00$ ($p = 0.955$).

All 139 were individual contributors, 20.84% of the 667, and no engagement manager or portfolio director wrote that the financial institution had not been told. The pattern was described by 21.74% of respondents on canceled engagements, against 12.62% to 13.20%

across the active stages. It was described by 21.55% (61 of 283) of respondents with a score of 3 or below, 14.77% (78 of 528) above 3 to 7, and none of the 39 above 7, and by 23.64% in the lowest indicator quintile against 10.00% in the highest. After Holm adjustment it differed by respondent role, the respondent's understanding of the obligations, the count of identified manual alteration methods, disclosure to the financial institution, and intent to leave.

The 178 answers came from all written answers to the open questions, and 94 of them were written on canceled engagements. Respondents described a team lead who knew of the manual alteration and a financial institution that did not, and one gave the financial institution's likely reaction as the reason.

> *“it's every single day now. i tell my manager, not the client, they'd panic. i'm just tired of making the call myself with no backup.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Corporate and back-office functions

> *“i export data manually and match it by hand because nothing come through api, team know, client does not know, this eat my whole morning some day”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Trade, settlement and reconciliation

Part 5 (Disclosure of Manual Alteration, section 2.5) analyzes the decision not to tell the financial institution within the answers on the three counted methods.

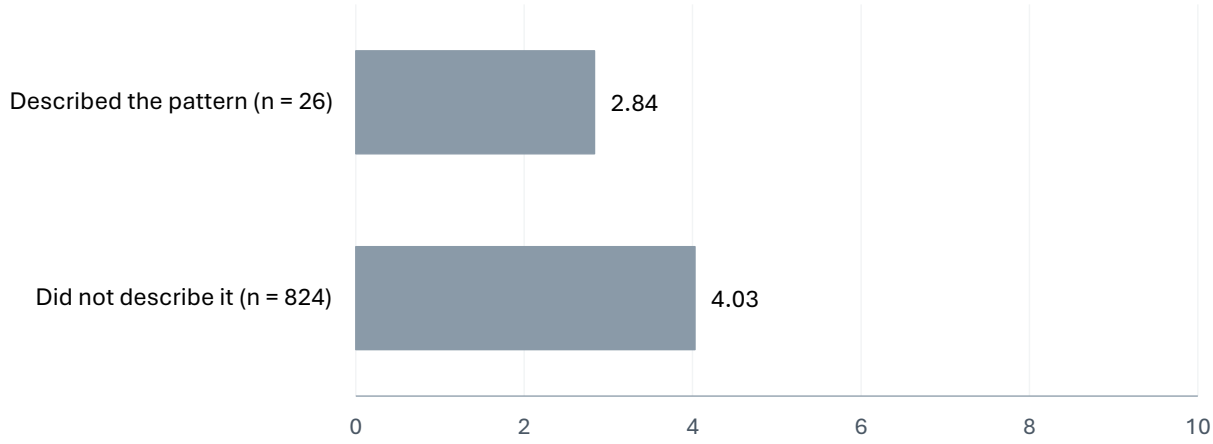
3.6 Financial Institution Told

Written pattern · All written answers to the open questions · 31 answers from 26 respondents on 23 engagements · Analyzed in Part 5, Disclosure of Manual Alteration, section 2.4

Accounts explicitly stating that the financial institution was told came from a small group with low mean scores. This finding does not establish that disclosure worsens governance or is ineffective. It may describe disclosure occurring amid existing operational problems, but the study does not establish that sequence. Its value is to distinguish confirmed reported telling from explicit nondisclosure and unknown disclosure.

Figure 3.121. Financial Institution Told: descriptive mean score comparison.

Accounts explicitly stating that the financial institution was told came from a small group with low mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 26 respondents who described it, a small base, recorded a score of 2.84, against 4.03 among the other 824 (cluster-robust $p < 0.001$). Among individual contributors alone, the 12 who described it stood at 3.17, against 3.88 among the other 655 (cluster-robust $p = 0.061$).

At the engagement level the share describing the pattern was associated with the score at $r = -0.30$ ($p < 0.001$, $n = 150$ engagements), and after adjustment for the mean of the other 53 indicators at partial $r = -0.06$ ($p = 0.482$).

Telling the financial institution rose with seniority, from 1.80% (12 of 667) of individual contributors to 5.33% (8 of 150) of engagement managers and 19.35% (6 of 31) of portfolio directors, a difference by role that held after Holm adjustment. The pattern was described by 4.97% of respondents on canceled engagements, against 0.97% to 2.40% across the active

stages, and by 15 of the 283 respondents with a score of 3 or below, 11 of the 528 above 3 to 7, and none of the 39 above 7.

The 31 answers came from all written answers to the open questions, and 20 of them were written on canceled engagements. Telling did not stop the manual alteration in the accounts of those who described it. An engagement manager wrote that the team flagged the manual alteration to the financial institution and went on doing it, and a portfolio director on another canceled engagement described each process change routed through the joint change-control forum.

> *"What we did, throughout this engagement, was route every process change through the joint change-control forum — so, nothing was altered without the client's calculation-methodology owner present. Attribution exceptions were raised, discussed and signed off before any workaround was applied; on top of that, that discipline held even as the account came under pressure. We do not deviate from the agreed process — on that basis the low frequency, six occasions in total, all minuted and shared with the client."*
>
> **PORTFOLIO DIRECTOR**
>
> KPO, Trade, settlement and reconciliation

Part 5 (Disclosure of Manual Alteration, section 2.4) analyzes the answers stating that the financial institution knew of the manual alteration within the answers on the three counted methods.

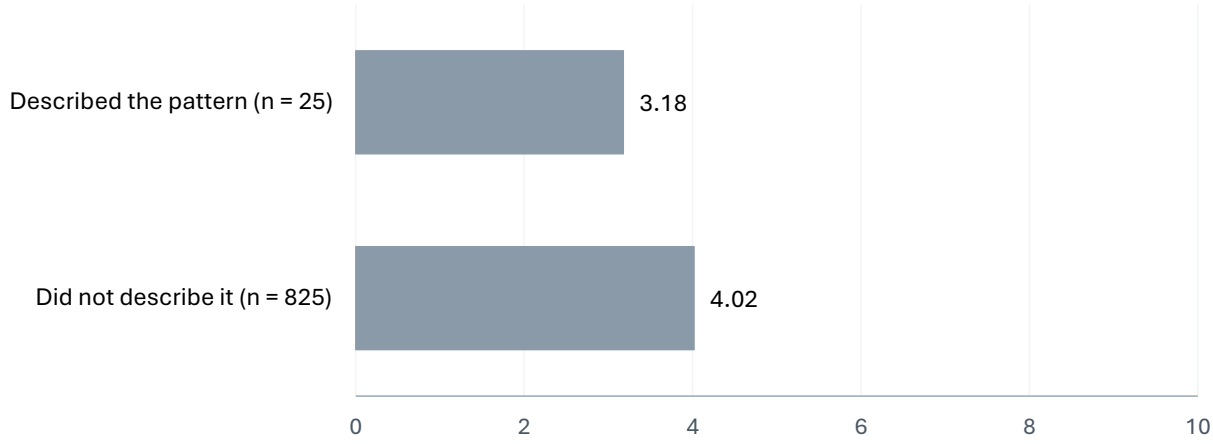
3.7 Disclosure Unknown to the Respondent

Written pattern · All written answers to the open questions · 30 answers from 25 respondents on 20 engagements · Analyzed in Part 5, Disclosure of Manual Alteration, section 2.6

Uncertainty about whether the financial institution had been told was associated with lower mean scores in the small group describing it. This identifies a visibility limit within the delivery workforce, rather than confirmed concealment or confirmed disclosure. The distinction matters because staff performing manual work may not know how it is communicated elsewhere. The analysis cannot resolve that uncertainty from nonmention.

Figure 3.122. Disclosure Unknown to the Respondent: descriptive mean score comparison.

Uncertainty about whether the financial institution had been told was associated with lower mean scores in the small group describing it.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 25 respondents who described the uncertainty, a small base, recorded a score of 3.18, against 4.02 among the other 825 (cluster-robust p = 0.003). All 25 were individual contributors, and against the other 642 individual contributors they stood at 3.18 and 3.89 (cluster-robust p = 0.014).

At the engagement level the share describing the pattern was associated with the score at r = −0.20 (p = 0.015, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.03 (p = 0.728).

The pattern was described by 5.28% of respondents on canceled engagements, against 1.20% to 1.94% across the active stages, and by 6.67% in the lowest indicator quintile against 1.25% in the highest. It was described by 12 of the 283 respondents with a score of 3 or below, 13 of

the 528 above 3 to 7, and none of the 39 above 7. After Holm adjustment it differed only by the count of identified manual alteration methods and by disclosure to the financial institution.

The 30 answers came from all written answers to the open questions, and 20 of them were written on canceled engagements. Respondents described breaks cleared without anyone above asking how, and steps adjusted without waiting for the ticket.

> *“We just patch the breaks ourselves and keep the queue moving, nobody upstairs asks how they got cleared. And I couldn't tell you if the client ever sees the real touch count.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Trade, settlement and reconciliation

> *“getting anything through formal change takes forever so everyone just works around it, we adjust the steps and dont wait for the ticket, dont know if client hears.”*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Payments, cards and disputes

Part 5 (Disclosure of Manual Alteration, section 2.6) analyzes the uncertainty over whether the financial institution had been told.

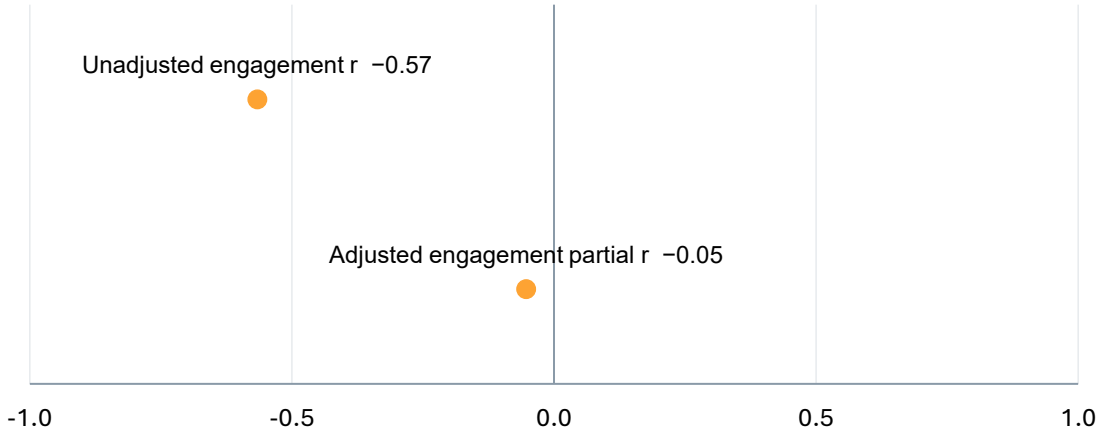
3.8 Problems Raised With No Response

Written pattern · All written answers to the open questions · 326 answers from 255 respondents on 119 engagements · Analyzed in Part 5, Manual Alteration by the Delivery Team, "Reasons Given for Manual Alteration"

Problems raised without a response were associated with low scores in the observed comparisons. Their governance significance lies in the gap between recognizing an issue and obtaining action from those responsible. The adjusted engagement association was much weaker. That model result limits a distinct-association claim without showing that unanswered escalation has no operational consequence.

Figure 3.123. Problems Raised With No Response: unadjusted and adjusted engagement associations with the score.

Problems raised without a response were associated with low scores in the observed comparisons.



Comparison	Estimate and analysis base
Unadjusted engagement r	r = −0.57 150 engagements
Adjusted engagement partial r	partial r = −0.05 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

The adjustment uses the mean of the other 53 indicators. Its p value is nominal. Written identification does not establish incidence, and the relationship is observational.

The 255 respondents who described it recorded a score of 3.17, against 4.35 among the other 595 (cluster-robust $p < 0.001$). Among individual contributors alone, the 220 who described it stood at 3.13, against 4.23 among the other 447 (cluster-robust $p < 0.001$).

At the engagement level the share describing the pattern was associated with the score at r = −0.57 (95% CI −0.67 to −0.45, p < 0.001, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.05 (p = 0.520).

The pattern was described by 47.00% (133 of 283) of respondents with a score of 3 or below, 22.92% (121 of 528) above 3 to 7, and 1 of the 39 above 7. It was described by 44.10% of respondents on canceled engagements, against 17.48% to 25.14% across the active stages, and by 46.67% in the lowest indicator quintile against 11.88% in the highest. Individual contributors described it at 32.98% (220 of 667), engagement managers at 19.33% (29 of 150), and portfolio directors at 19.35% (6 of 31), a difference by role that held after Holm adjustment. The pattern differed after Holm adjustment in 11 of the 19 segment families, and in none of the structural families of institution type, outsourcing model, process family, class of work, vendor type, and delivery location.

The 326 answers came from all written answers to the open questions, and 195 of them were written on canceled engagements. Respondents described a request passed from a contributor to a team lead to the financial institution without reply, and flags left to sit in a queue checked the next day.

> *“to be blunt i was angry. we ask for one contact who can approve a new supplier record, for three week nobody reply, i follow up with my TL, he follow up with them, still nothing. we also cannot see their supplier history so we working blind”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Corporate and back-office functions

> *“the problem is nobody watches the agents when they misfile a scan. we flag it and it just sits. what would actually fix it is one person on our side owning exceptions for the shift, instead of a queue somebody checks tommorow. training i got was fine, that part i can't complain about”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Corporate and back-office functions

Part 5 (Manual Alteration by the Delivery Team, “Reasons Given for Manual Alteration”) analyzes problems raised with no response within the answers on the three counted methods.

3.9 Raising Problems Seen as Futile or Discouraged

Written pattern · All written answers to the open questions · 149 answers from 122 respondents on 81 engagements · Analyzed in Part 5, Manual Alteration by the Delivery Team, "Reasons Given for Manual Alteration"

Seeing escalation as futile or discouraged was associated with low scores, including a negative adjusted estimate. This concerns a reported barrier to making problems visible, distinct from a problem raised without response or a direct instruction not to raise it. The interpretation remains conditional on the specified model and test family. It does not independently establish the effect of discouragement on engagement outcomes.

Figure 3.124. Raising Problems Seen as Futile or Discouraged: unadjusted and adjusted engagement associations with the score.

Seeing escalation as futile or discouraged was associated with low scores, including a negative adjusted estimate.

Comparison	Estimate and analysis base
Unadjusted engagement r	r = −0.49 150 engagements
Adjusted engagement partial r	partial r = −0.22 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

The adjustment uses the mean of the other 53 indicators. Its p value is nominal. Written identification does not establish incidence, and the relationship is observational.

The 122 respondents who described it recorded a score of 2.69, against 4.21 among the other 728 (cluster-robust $p < 0.001$). Among individual contributors alone, the 119 who described it stood at 2.71, against 4.11 among the other 548 (cluster-robust $p < 0.001$).

At the engagement level the share describing the pattern was associated with the score at r = −0.49 (95% CI −0.60 to −0.35, p < 0.001, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.22 (p = 0.008).

The pattern was described by 27.92% (79 of 283) of respondents with a score of 3 or below, 8.14% (43 of 528) above 3 to 7, and none of the 39 above 7. It was described by 20.81% of respondents on canceled engagements, against 10.29% to 10.68% across the active stages, and by 24.24% in the lowest indicator quintile against 3.75% in the highest. Individual contributors described it at 17.84% (119 of 667), and 3 of the 150 engagement managers and none of the 31 portfolio directors did, a difference by role that held after Holm adjustment. The pattern differed after Holm adjustment in 11 of the 19 segment families and in none of the structural families.

The 149 answers came from all written answers to the open questions, and 86 of them were written on canceled engagements. Respondents described queries answered after the case had moved on, colleagues who then decided alone and wrote a note, and a team that stopped raising disputes left open for days.

> *“The pace is what wear on my team. We raise a query on an escalation and the answer come three days later, by then the case has moved on anyway. So people stop asking. Two of my colleagues just decide and write a note now, and I am doing same, though nobody has asked us to. And the quality checks they run on us do not look at the errors we actually make, so nobody above sees any of this”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Financial crime and client onboarding

> *“I stop raising things. Two of us sit on the same dispute case for nine days waiting for someone to say yes or no, then it closes on its own because the customer gave up. That does something to you, i dont care about this project now, i just do my hours.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Payments, cards and disputes

Part 5 (Manual Alteration by the Delivery Team, “Reasons Given for Manual Alteration”) analyzes futile or discouraged escalation within the answers on the three counted methods.

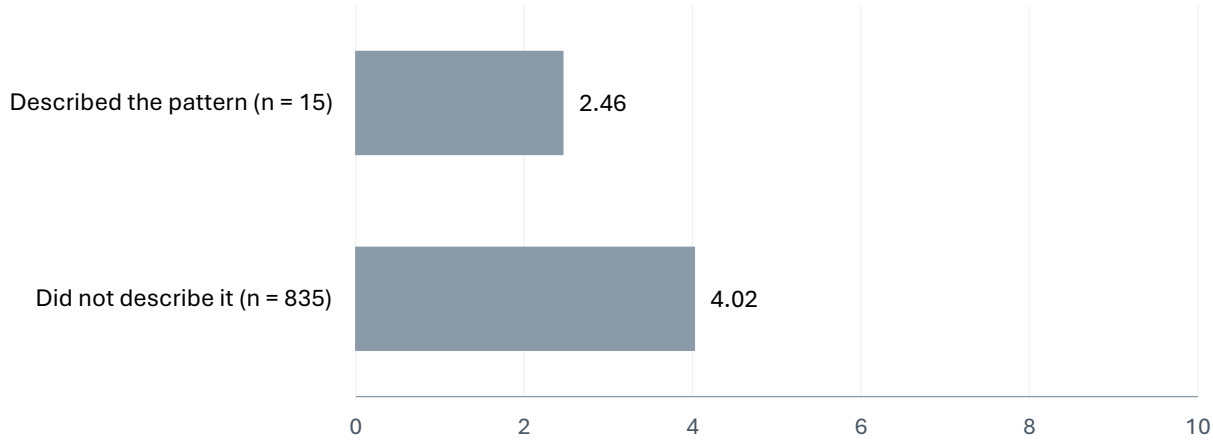
3.10 Instructed Not to Raise the Problem

Written pattern · All written answers to the open questions · 20 answers from 15 respondents on 12 engagements · Analyzed in Part 5, Manual Alteration by the Delivery Team, "Reasons Given for Manual Alteration"

Explicit instructions not to raise a problem were described by a small group with low mean scores. The engagement-level raw association was negative, while the adjusted analysis did not detect a clear association. The 12-engagement reach warrants caution but does not negate the reported raw result. An instruction not to raise a problem directly restricts escalation and is distinct from perceived futility or an unanswered request.

Figure 3.125. Instructed Not to Raise the Problem: descriptive mean score comparison.

Explicit instructions not to raise a problem were described by a small group with low mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

Fifteen respondents on 12 engagements described an instruction not to raise a problem. Their mean score was 2.46, compared with 4.02 among the other 835 (cluster-robust $p < 0.001$). Among individual contributors, the contrast was 2.65 for the 13 describing it and 3.89 for the other 654 (cluster-robust $p < 0.001$). The small base warrants caution. It does not negate the raw engagement association reported separately.

At the engagement level the share describing the pattern was associated with the score at $r = -0.24$ ($p = 0.003$, $n = 150$ engagements), and after adjustment for the mean of the other 53 indicators at partial $r = -0.04$ ($p = 0.588$).

The instruction was described by 12 of the 283 respondents with a score of 3 or below, 3 of the 528 above 3 to 7, and none of the 39 above 7. Eleven of the 15 respondents were on canceled engagements, and 9 of the 12 engagements were canceled. Thirteen of the 15 were individual contributors and two were engagement managers, and no segment family differed after Holm adjustment.

The 20 answers came from all written answers to the open questions, and 16 of them were written on canceled engagements. The pattern is the written evidence of an instruction from above not to raise the problem. Respondents described an instruction to keep the numbers looking steady without explaining their movement, and a manager who treated a flag as a sign that the respondent could not manage the workload.

> *Eight times out of ten someone on my team is fixing the output before it goes anywhere, catching bad attribution numbers before a client sees them... we are told to keep the numbers looking steady, not to explain why they moved."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> KPO, Trade, settlement and reconciliation

> *the problem is if i flag this stuff my Manager treats it like i cant handle the workload, so i fix it and move on, the exceptions clear either way"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Payments, cards and disputes

Part 5 (Manual Alteration by the Delivery Team, "Reasons Given for Manual Alteration") analyzes the instruction not to raise the problem within the answers on the three counted methods.

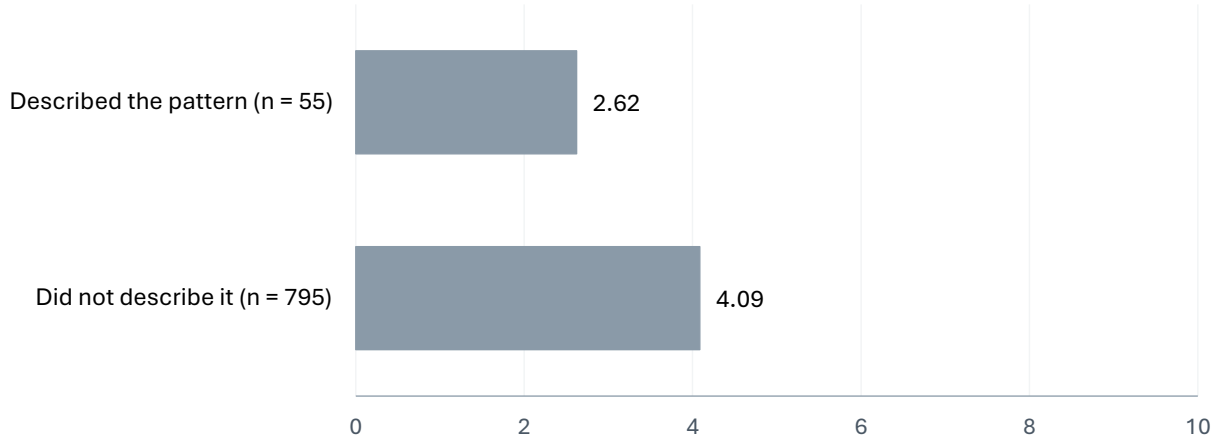
3.11 Manual Alteration Disclosed to No Party

Written pattern · All written answers to the open questions · 59 answers from 55 respondents on 39 engagements · Analyzed in Part 5, Disclosure of Manual Alteration, section 3.1

Manual alteration disclosed to no party was associated with low scores and was described entirely by individual contributors. This concerns operational work that remains outside both internal management discussion and financial institution disclosure. The written accounts identify a specific visibility gap. They do not enumerate every unreported intervention or establish that nondisclosure caused the low scores.

Figure 3.126. Manual Alteration Disclosed to No Party: descriptive mean score comparison.

Manual alteration disclosed to no party was associated with low scores and was described entirely by individual contributors.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 55 respondents who described it recorded a score of 2.62, against 4.09 among the other 795 (cluster-robust $p < 0.001$). All 55 were individual contributors, and against the other 612 individual contributors they stood at 2.62 and 3.98 (cluster-robust $p < 0.001$).

At the engagement level the share describing the pattern was associated with the score at $r = -0.30$ ($p < 0.001$, $n = 150$ engagements), and after adjustment for the mean of the other 53 indicators at partial $r = -0.10$ ($p = 0.235$).

Thirty-nine of the 55 respondents were on canceled engagements, where 12.11% described the pattern, against 1.71% to 4.00% across the active stages, and 26 of the 39 engagements were canceled. The pattern was described by 13.43% (38 of 283) of respondents with a score of 3 or below, 3.22% (17 of 528) above 3 to 7, and none of the 39 above 7, and by 11.52% in the

lowest indicator quintile against 1.25% in the highest. After Holm adjustment it differed by indicator quintile, engagement status, cancellation risk band, the financial institution's discussion of the obligations, regulatory training, the count of identified manual alteration methods, disclosure to the financial institution, and intent to leave.

The 59 answers came from all written answers to the open questions, and 41 of them were written on canceled engagements. Respondents described corrections known to neither their manager nor the financial institution, and silence that followed a problem raised once without result.

> *Nobody know I do this, not my manager and not the client. I just fix and move on because otherwise report is late every time"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Regulatory, audit, tax and legal

> *Nobody know I do this, not manager, not anyone, I dont say anything anymore because last time I raise it nothing happen, so now I just fix and stay quiet"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Client servicing and communications

Part 5 (Disclosure of Manual Alteration, section 3.1) analyzes manual alteration disclosed to no party within the answers on the three counted methods.

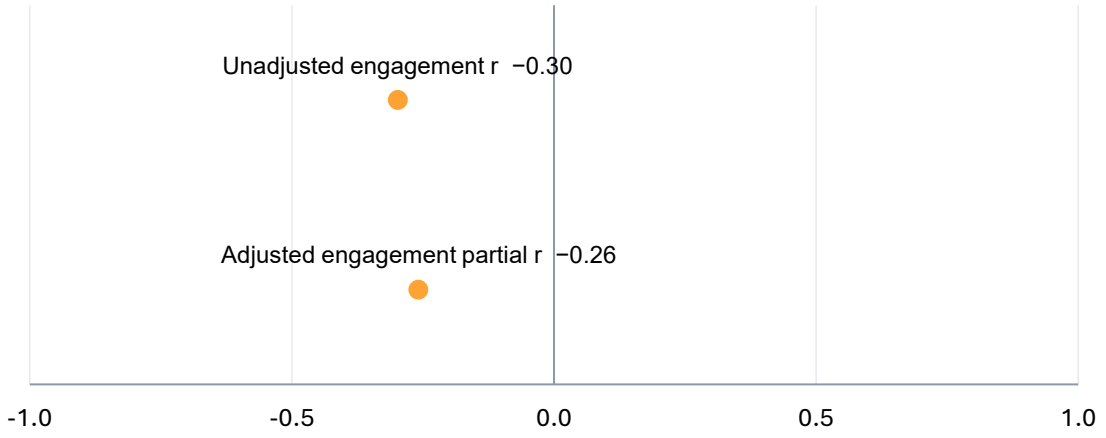
3.12 Manual Alteration Not Logged or Only Partly Logged

Written pattern · All written answers to the open questions · 143 answers from 118 respondents on 72 engagements · Analyzed in Part 5, Disclosure of Manual Alteration, section 3.2

Unlogged or partly logged manual alteration retained a negative association with the score under the main adjustment. This directly concerns the evidence available for oversight: work cannot be fully reconstructed where its manual component is absent from the record. The status distribution supplies context, while the adjusted association answers the main governance question. Both depend on written identification rather than a comprehensive log audit.

Figure 3.127. Manual Alteration Not Logged or Only Partly Logged: unadjusted and adjusted engagement associations with the score.

Unlogged or partly logged manual alteration retained a negative association with the score under the main adjustment.



Comparison	Estimate and analysis base
Unadjusted engagement r	r = −0.30 150 engagements
Adjusted engagement partial r	partial r = −0.26 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

The adjustment uses the mean of the other 53 indicators. Its p value is nominal. Written identification does not establish incidence, and the relationship is observational.

The 118 respondents who described it recorded a score of 3.14, against 4.13 among the other 732 (cluster-robust $p < 0.001$). Among individual contributors alone, the 109 who described it stood at 3.23, against 3.99 among the other 558 (cluster-robust $p < 0.001$).

At the engagement level the share describing the pattern was associated with the score at $r = -0.30$ (95% CI −0.44 to −0.14, $p < 0.001$, $n = 150$ engagements). After adjustment for the mean of the other 53 indicators the adjusted estimate was partial $r = -0.26$ ($p = 0.001$), the strongest adjusted association of the 41 written patterns tested on all written answers.

The pattern was described by 18.45% of respondents in Production and 17.70% on canceled engagements, against 8.00% in Build-Out and 11.20% in Transition/Cut-Over. It was described by 21.91% (62 of 283) of respondents with a score of 3 or below, 10.61% (56 of 528) above 3 to 7, and none of the 39 above 7, and by 23.64% in the lowest indicator quintile against 8.12% in the highest. Individual contributors described it at 16.34% (109 of 667), engagement managers at 6.00% (9 of 150), and no portfolio director did, a difference by role that held after Holm adjustment.

The 143 answers came from all written answers to the open questions, and 69 of them were written on canceled engagements. Respondents described half of their fixes missing from any log, and mismatches reconciled with no note of which system had been right.

> *Half of what we fix never make it into any log... and another thing nobody above us even ask about it anymore, so we just decide and move on with the fix"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Corporate and back-office functions

> *my team hits this seven times out of ten, we just reconcile the mismatches ourselves and dont log which system was right, theres no time to write it up"*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Payments, cards and disputes

Part 5 (Disclosure of Manual Alteration, section 3.2) analyzes manual alteration not logged or only partly logged within the answers on the three counted methods.

“

On my last account we wait three weeks for one AI person to be assigned, here I get a specialist on the same day I ask. The training also is better now, they teach us the payments side properly. Only thing, if an agent makes a costly mistake I am not sure who stands up for it, we never discuss about that”

INDIVIDUAL CONTRIBUTOR

CXM, Corporate and back-office functions

4 Additional Reporting Practices Below the Threshold for Statistical Analysis

Three additional reporting practices identify control questions through very small written bases. They concern attribution of human work, the elapsed time assigned to rework, and the age of unresolved items. Each remains a separate descriptive finding. The paper's minimum count for comparative analysis is not met, so the score contrasts are descriptions without inferential tests.

4.1 Human Work Presented as Agentic

Written pattern · All written answers to the open questions · 6 answers from 6 respondents on 5 engagements · Listed in Part 5, Additional Conditions Below the Threshold for Statistical Analysis

Six respondents described human work being presented as agentic. The governance relevance is the attribution of delivery work and the resulting visibility of human intervention. The accounts are too few for the paper's comparative statistical analysis.

Figure 3.128. Human Work Presented as Agentic: descriptive evidence and its limits

Six respondents described human work being presented as agentic.

Reported practice	Governance implication	Evidence boundary
Six respondents described human work being presented as agentic.	The governance relevance is the attribution of delivery work and the resulting visibility of human intervention.	The accounts are descriptive. They do not establish the prevalence of the practice or a causal effect.

Note: Respondents describing human Work Presented as Agentic

Small base. No percentage comparison between status groups.

6 written accounts from 6 respondents on 5 engagements.

Mean score 3.15 among those describing the pattern, compared with 4.00 among the remainder. Descriptive comparison only, with no inferential test.

The six respondents recorded a score of 3.15, against 4.00 among the rest, a description without a test. Five of the six were on canceled engagements, and they were three individual contributors, two engagement managers, and one portfolio director.

A portfolio director on a canceled engagement in regulatory, audit, tax and legal work described the delivery model inverting within six weeks, engineers writing the logic and the agent producing the documentation, with the change never recorded as a design change.

> *"What we ended up doing — and I would want this recorded, because I think it is the actual finding of the engagement — is that we staffed around the tool. The design had the agent drafting surveillance logic with an engineer reviewing it. Within about six weeks that inverted. The engineers wrote the scenario logic themselves and used the agent for the documentation and the test cases around it, because the documentation is where it was strong and the logic is where it kept losing the thread. It would hold a constraint through the first half of a long session and then quietly stop applying it, so anything needing more than an hour of back and forth had to be broken into pieces small enough that nothing needed carrying between them. We never wrote that down as a design change. It happened at desk level and I allowed it, because the alternative was telling the client we had inverted the delivery model in month two. I would handle that differently now. The pattern is going to repeat on every regulatory build we take on, and if we do not name it we will keep selling work as agent-led that is in practice engineer-led with an agent doing the paperwork."*
>
> **PORTFOLIO DIRECTOR**
>
> ITO, Regulatory, audit, tax and legal

Part 5 (Additional Conditions Below the Threshold for Statistical Analysis) lists human work presented as agentic among the answers on manual alteration.

4.2 Elapsed-Time Measures Restarted on Rework

Written pattern · All written answers to the open questions · 4 answers from 4 respondents on 3 engagements · Analyzed in Part 5, Reporting alteration, section 3.5

Four respondents described restarting elapsed-time measurement when work was redone. This affects what a completion-time measure captures and can leave the history of rework outside the reported duration. The score comparison is descriptive and was not tested. The finding concerns the reported measurement practice, with no established difference between the small group and other respondents.

Figure 3.129. Elapsed-Time Measures Restarted on Rework: descriptive evidence and its limits

Four respondents described restarting elapsed-time measurement when work was redone.

Reported practice	Governance implication	Evidence boundary
Four respondents described restarting elapsed-time measurement when work was redone.	This affects what a completion-time measure captures and can leave the history of rework outside the reported duration.	The score comparison is descriptive and was not tested.

Note: Respondents describing elapsed-Time Measures Restarted on Rework

Small base. No percentage comparison between status groups.

4 written accounts from 4 respondents on 3 engagements.

Mean score 3.90 among those describing the pattern, compared with 3.99 among the remainder. Descriptive comparison only, with no inferential test.

The four respondents describing restarted elapsed-time measures had a mean score of 3.90, compared with 3.99 among the rest. No inferential test was performed, so the comparison does not establish either a difference or equivalence. Two accounts came from Build-Out, one from Transition/Cut-Over, and one from a canceled engagement.

An engagement manager described a misrouted ticket that received a fresh timestamp on reassignment, so that chronic misrouting never registered as delay.

> *The service management dashboard reports ticket resolution time from the moment the agent first touches a ticket, but any ticket the agent routes incorrectly and has to be manually reassigned gets a fresh timestamp on reassignment, so chronic misrouting never shows up as a delay in the headline metric."*
>
> **ENGAGEMENT MANAGER**
>
> CXM, Technology engineering and operations

Part 5 (Reporting alteration, section 3.5) reports the practice within the answers on the manual alteration of reporting.

4.3 Suppressed Aging of Open Items

Written pattern · All written answers to the open questions · 3 answers from 3 respondents on 3 engagements · Listed in Part 5, Additional Conditions Below the Threshold for Statistical Analysis

Three respondents described suppression of the aging of open items. This concerns whether oversight can see how long unresolved work has remained outstanding, distinct from restarting a timer on rework. The small base supports descriptive reporting only. It does not establish prevalence, a score gradient, or the motives and consequences of each reported practice.

Figure 3.130. Suppressed Aging of Open Items: descriptive evidence and its limits

Three respondents described suppression of the aging of open items.

Reported practice	Governance implication	Evidence boundary
Three respondents described suppression of the aging of open items.	This concerns whether oversight can see how long unresolved work has remained outstanding, distinct from restarting a timer on rework.	The small base supports descriptive reporting only.

Note: Respondents describing suppressed Aging of Open Items

Small base. No percentage comparison between status groups.

3 written accounts from 3 respondents on 3 engagements.

Mean score 3.40 among those describing the pattern, compared with 3.99 among the remainder. Descriptive comparison only, with no inferential test.

The three respondents recorded a score of 3.40, against 3.99 among the rest, a description without a test. Two of them were in Production and one on a canceled engagement.

Respondents described custody breaks grouped under a pending status that never aged, and certification delays rolled into a single line on the dashboard.

> *The vendor's exception dashboard groups unresolved custody breaks under a catch-all pending category that does not age, so a break sitting open for weeks looks the same as one opened yesterday."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Trade, settlement and reconciliation

> *The vendor rolled certification delays into a single 'network dependency' line on the dashboard so the customer never saw which processor gate was actually stuck or for how long."*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Payments, cards and disputes

Part 5 (Additional Conditions Below the Threshold for Statistical Analysis) lists the practice among the answers on manual alteration.

> *Theres a model defect form youre supposed to file when the Bot gets a fix wrong. Nobody fills it out, takes about twenty minutes and it goes to a queue that answers in like three weeks. So my team keeps a notepad file in Teams with what actually worked on which ticket type, and we just use that."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

“

How I feel is caught in the middle, genuinely, because my team override the early-warning alerts before they escalate them, not for fun, but because the signal needs a human check before anyone trusts it. The decision on which alerts even matter sits with a forum that meets far less than the alerts fire, and what reaches the portfolio director is only what my analysts choose to tell me, far as I know.”

ENGAGEMENT MANAGER

KPO, Credit, underwriting and suitability

Practices Adopted in Place of Controls

Human rework and normalized workarounds were associated with lower governance, oversight, and control scores, while habit in place of an explicit control showed no clearly detected gradient. These practices connect delivery continuity with work that formal control and reporting may represent incompletely. Their differing relationships with the composite matter: a specific unwritten dependency can be reported even where the broader score is relatively favorable.

These practices distinguish correction of agent work from its visibility in the record used for oversight. Human rework and normalized workarounds had lower observed scores, while unreported effort showed a weaker adjusted relationship and habit showed no clearly detected gradient. Four of the 39 respondents above score 7 described habit in place of a control. The normalized-workaround partial correlation of −0.18 is nominally significant and remains subject to the stated test-family qualification. Table 3.18 retains each practice and the separately identified proposed remedies.

Table 3.18. The four practices adopted in place of controls and the remedies proposed by respondents against the score

Written pattern	Base	Engagements	The score, describing against the rest (cluster-robust p)
Human rework absorbing agent failure	344 answers from 244 respondents	111	2.88 against 4.44 (p < 0.001)
Workarounds normalized into the process	152 answers from 137 respondents	86	3.16 against 4.15 (p < 0.001)
Effort absent from reported figures	71 answers from 64 respondents	54	3.14 against 4.06 (p < 0.001)
Habit in place of a control	91 answers from 87 respondents	69	3.93 against 4.00 (p = 0.705)
Remedies proposed by respondents	225 answers from 203 respondents	118	4.05 against 3.98 (p = 0.600)

Written pattern	Engagement r (p)	Adjusted partial r (p)
Human rework absorbing agent failure	−0.57 (p < 0.001)	−0.11 (p = 0.180)
Workarounds normalized into the process	−0.36 (p < 0.001)	−0.18 (p = 0.027)
Effort absent from reported figures	−0.27 (p < 0.001)	−0.14 (p = 0.098)
Habit in place of a control	−0.02 (p = 0.789)	−0.01 (p = 0.872)
Remedies proposed by respondents	−0.01 (p = 0.868)	−0.11 (p = 0.194)

Note: The four practices and proposed remedies use written answers to the open questions. Correlation r relates the engagement's share describing a practice to its score, across 150 engagements. Partial r accounts for the mean of the 53 indicators outside the score. The reported partial-correlation p values are nominal. Description of a practice does not establish its frequency, and nonmention does not establish its absence.

1 Workarounds Adopted as Standard Practice

The four practices distinguish human correction, routine workarounds, unreported effort, and dependence on habit. Rework and missing effort had negative observed relationships with the score, while normalized workarounds retained a nominal adjusted association. Habit showed no clearly detected relationship. Each contributes a different view of how actual work can diverge from the process visible to oversight.

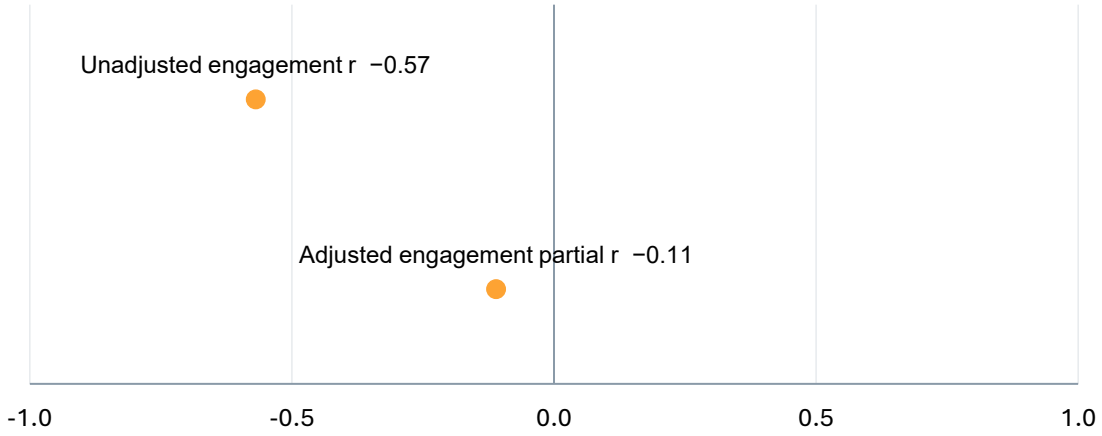
1.1 Human Rework Absorbing Agent Failure

Written pattern · All written answers to the open questions · 344 answers from 244 respondents on 111 engagements · Analyzed in Part 5, Manual Alteration of AI Outputs, section 2.1

Human rework absorbing agent failure was associated with lower mean scores and a strong negative observed engagement association. Operational delivery can depend on human correction whose contribution is incompletely represented in formal control and reporting. The adjusted association was weaker and not clearly detected. This distinguishes substantial reported rework from evidence of an independently identified governance effect.

Figure 3.131. Human Rework Absorbing Agent Failure: unadjusted and adjusted engagement associations with the score.

Human rework absorbing agent failure was associated with lower mean scores and a strong negative observed engagement association.



Comparison	Estimate and analysis base
Unadjusted engagement r	r = −0.57 150 engagements
Adjusted engagement partial r	partial r = −0.11 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

The adjustment uses the mean of the other 53 indicators. Its p value is nominal. Written identification does not establish incidence, and the relationship is observational.

The 244 respondents who described it recorded a score of 2.88, against 4.44 among the other 606 (cluster-robust $p < 0.001$). Among individual contributors alone, the 232 who described it stood at 2.87, against 4.40 among the other 435 (cluster-robust $p < 0.001$).

At the engagement level the share describing the practice was associated with the score at r = −0.57 (95% CI −0.67 to −0.45, p < 0.001, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.11 (p = 0.180).

The practice was described by 52.30% (148 of 283) of respondents with a score of 3 or below, 17.80% (94 of 528) above 3 to 7, and 2 of the 39 above 7. It was described by 44.41% of respondents on canceled engagements, against 16.00% in Build-Out, 19.60% in Transition/Cut-Over, and 23.30% in Production, and by 51.52% in the lowest indicator quintile against 10.00% in the highest. Individual contributors described it at 34.78% (232 of 667), engagement managers at 7.33% (11 of 150), and 1 of the 31 portfolio directors, a difference by role that held after Holm adjustment. The practice differed after Holm adjustment in 11 of the 19 segment families and in none of the structural families.

Rework followed the capacity of human review. Among respondents who rated human review capacity against agent volume 1 to 3, 43.57% (166 of 381) described rework, against 14.90% (59 of 396) of those who rated it 4 to 7 and 1 of the 28 who rated it 8 to 10 (chi-square = 87.67, p < 0.001, Cramér's V = 0.33).

The 344 answers came from all written answers to the open questions, and 207 of them were written on canceled engagements. Respondents described a funding cutoff missed twice and reconciled manually while a rule change waited for approval, and documents checked manually where no test environment had proved the agent before live work began.

> *“there is no test setup, we could not prove anything before it ran on live cases. passport and address proof it reads sometimes, everything else i check by hand”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Financial crime and client onboarding

Part 5 (Manual Alteration of AI Outputs, section 2.1) analyzes human rework within the answers on the manual alteration of AI outputs.

1.2 Workarounds Normalized Into the Process

Written pattern · All written answers to the open questions · 152 answers from 137 respondents on 86 engagements · Analyzed in Part 5, Manual Alteration of Financial Institution Processes, section 2.4

Workarounds normalized into the process were associated with lower scores and a negative adjusted estimate. When an exceptional correction becomes routine, its visibility and formal control may not develop with it. The nominal adjusted result requires the stated multiple-testing qualification. It supports retaining the pattern as a distinct analytical subject without treating its significance as settled under every test family.

Figure 3.132. Workarounds Normalized Into the Process: unadjusted and adjusted engagement associations with the score.

Workarounds normalized into the process were associated with lower scores and a negative adjusted estimate.

Comparison	Estimate and analysis base
Unadjusted engagement r	r = −0.36 150 engagements
Adjusted engagement partial r	partial r = −0.18 150 engagements

Note: r measures signed linear association from −1 to +1. Partial r describes the adjusted association under the model specified in the section. Adjusted p values are nominal. The comparisons do not establish causal effects.

The adjustment uses the mean of the other 53 indicators. Its p value is nominal. Written identification does not establish incidence, and the relationship is observational.

The 137 respondents who described them recorded a score of 3.16, against 4.15 among the other 713 (cluster-robust $p < 0.001$). Among individual contributors alone, the 130 who described them stood at 3.15, against 4.04 among the other 537 (cluster-robust $p < 0.001$).

At the engagement level the share describing the practice was associated with the score at r = −0.36 (95% CI −0.50 to −0.22, p < 0.001, n = 150 engagements). After adjustment for the mean of the other 53 indicators the adjusted estimate was partial r = −0.18 (p = 0.027), one of three written patterns of the 41 to keep an adjusted association.

The practice was described by 24.24% of respondents in the lowest indicator quintile and 5.00% in the highest. It was described by 25.44% (72 of 283) of respondents with a score of 3 or below, 12.31% (65 of 528) above 3 to 7, and none of the 39 above 7. It was described by 21.43% of respondents on canceled engagements, against 12.57% to 13.59% across the active stages, a difference that did not survive Holm adjustment. Individual contributors described it at 19.49% (130 of 667), engagement managers at 4.67% (7 of 150), and no portfolio director did, a difference by role that held after Holm adjustment.

The 152 answers came from all written answers to the open questions, and 75 of them were written on canceled engagements. Respondents described workarounds that were never signed off and that became normal with time, and silence from the financial institution taken as approval.

> *“my team does this most days now, in addition nobody signs off the workaround, it just becomes normal after time”*
>
> **INDIVIDUAL CONTRIBUTOR**
> BPO, Trade, settlement and reconciliation

> *“This happens every week, when i raise it with Client they just revert back with nothing useful, so silence is treated as approval by us now”*
>
> **INDIVIDUAL CONTRIBUTOR**
> KPO, Trade, settlement and reconciliation

Part 5 (Manual Alteration of Financial Institution Processes, section 2.4) analyzes the workaround within the answers on the manual alteration of processes.

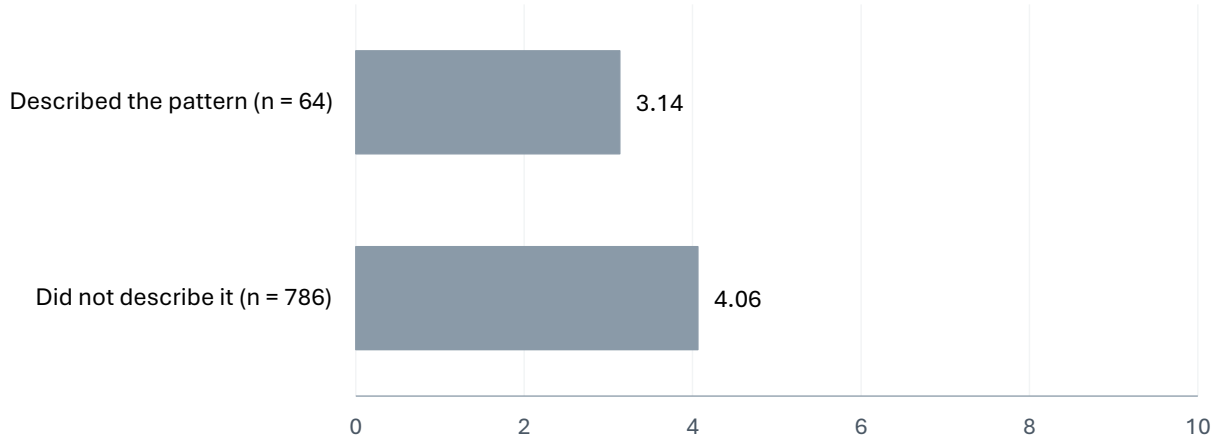
1.3 Effort Absent From Reported Figures

Written pattern · All written answers to the open questions · 71 answers from 64 respondents on 54 engagements · Analyzed in Part 5, Disclosure of Manual Alteration, section 3.3

Effort absent from reported figures was described by respondents with lower mean scores. This connects the record used for oversight with the work required to sustain delivery. The engagement association was negative, while the adjusted analysis did not detect a clear separate association. The reported visibility gap remains substantively relevant, while the adjusted result limits a claim of a separate statistical association.

Figure 3.133. Effort Absent From Reported Figures: descriptive mean score comparison.

Effort absent from reported figures was described by respondents with lower mean scores.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 64 respondents who described it recorded a score of 3.14, against 4.06 among the other 786 (cluster-robust $p < 0.001$). Among individual contributors alone, the 50 who described it stood at 3.07, against 3.93 among the other 617 (cluster-robust $p = 0.002$).

At the engagement level the share describing the practice was associated with the score at $r = -0.27$ (95% CI −0.41 to −0.11, $p < 0.001$, $n = 150$ engagements), and after adjustment for the mean of the other 53 indicators at partial $r = -0.14$ ($p = 0.098$).

The practice differed across engagement status, described by 2.86% of respondents in Build-Out, 6.00% in Transition/Cut-Over, 7.77% in Production, and 11.18% on canceled engagements. It was described by 11.31% (32 of 283) of respondents with a score of 3 or below, 6.06% (32 of 528) above 3 to 7, and none of the 39 above 7. Engagement managers

described it at 9.33% (14 of 150) and individual contributors at 7.50% (50 of 667), with no clearly detected difference by role after Holm adjustment. After Holm adjustment it differed by indicator quintile, engagement status, cancellation risk band, the financial institution's discussion of the obligations, process family, the count of identified manual alteration methods, and disclosure to the financial institution.

The 71 answers came from all written answers to the open questions, and 41 of them were written on canceled engagements. An engagement manager described two nights of manual re-tagging on work no count recorded, and an individual contributor described errors caught before release that appeared in no count.

> *“problem is nobody counts the wrong answers we caught, only ones the client complains about”*
>
> **INDIVIDUAL CONTRIBUTOR**
> CXM, Client servicing and communications

Part 5 (Disclosure of Manual Alteration, section 3.3) analyzes effort left out of the reported numbers within the answers on the three counted methods.

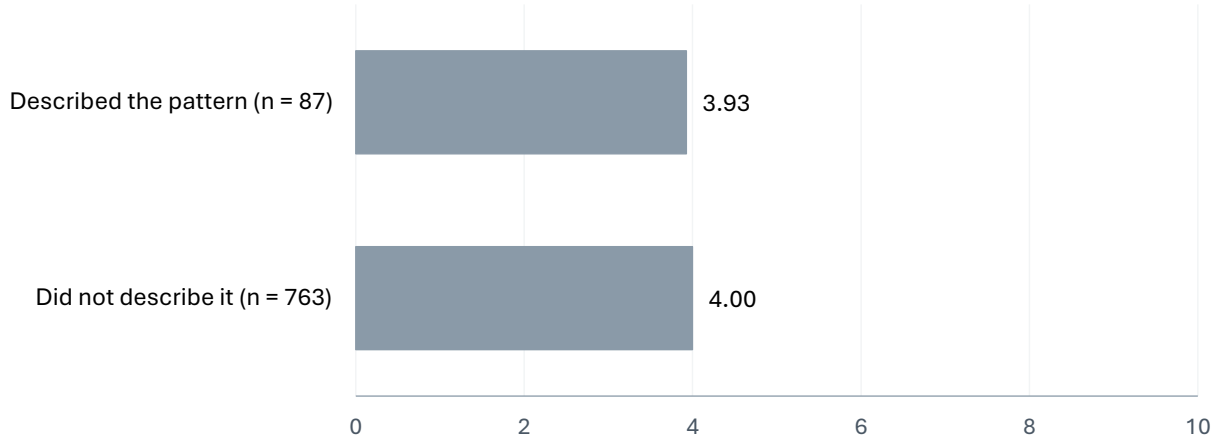
1.4 Habit in Place of a Control

Written pattern · All written answers to the open questions · 91 answers from 87 respondents on 69 engagements · Analyzed in Part 5, Disclosure of Manual Alteration, section 3.4

Habit standing in for an explicit control was described across score bands without a clearly detected score gradient. This is a distinct governance finding because favorable composite ratings can coexist with work that depends on particular people and unwritten routines. The analysis does not formally establish that the score fails as a detection instrument. It shows that this reported practice was not sharply distinguished by the measured comparisons.

Figure 3.134. Habit in Place of a Control: descriptive mean score comparison.

Habit standing in for an explicit control was described across score bands without a clearly detected score gradient.



Note: The comparison group did not describe the pattern and is not a verified nonincident group. The eligible source and inferential qualifications are stated in the text.

The 87 respondents who described it recorded a score of 3.93, against 4.00 among the other 763 (cluster-robust p = 0.705). Among individual contributors alone, the 64 who described it stood at 3.65, against 3.89 among the other 603 (cluster-robust p = 0.297).

At the engagement level the share describing the practice was associated with the score at r = −0.02 (p = 0.789, n = 150 engagements), and after adjustment for the mean of the other 53 indicators at partial r = −0.01 (p = 0.872).

The practice was described by 11.31% (32 of 283) of respondents with a score of 3 or below, 9.66% (51 of 528) above 3 to 7, and 4 of the 39 above 7. It was described by 11.52% in the lowest indicator quintile and 8.75% in the highest, and by 7.77% to 12.40% across engagement status. Portfolio directors described it at 19.35% (6 of 31), engagement managers at 11.33%

(17 of 150), and individual contributors at 9.60% (64 of 667), and no segment family differed after Holm adjustment.

The rating about replacing informal fixes with explicit checks did not show a clear association with the written habit pattern. The pattern was described by 27 of 267 respondents rating that item at 1–3, 47 of 459 at 4–7, and 10 of 70 at 8–10. The corresponding shares were 10.11%, 10.24%, and 14.29%. The ordinary chi-square result was 1.14 ($p = 0.567$, Cramér's $V = 0.04$), subject to the stated clustering limitation. This is not a formal validation test of the rating's detection capability.

The 91 answers came from all written answers to the open questions, and 36 of them were written on canceled engagements. An individual contributor who wrote that the providers gave notice of change described a colleague who read every release note unasked, a step written into no process. Another individual contributor described an engagement held together by two people talking directly, an arrangement that stopped when one of them was away for a week.

> *“Feel fine about this part mostly... the providers do tell us in advance when something change, that works. What these questions leave out is the person side. One colleague reads every release note from the model provider and posts a summary in our channel. He started it himself, nobody asked him to. It is not written into any process anywhere.”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Technology engineering and operations

> *“what actually held this together was my lead and one guy on their side talking directly, everything else on paper never got used... that is not a process, when my lead was out for a week nothing moved at all. none of these questions get at how much of it was just two people knowing each other”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> CXM, Client servicing and communications

Part 5 (Disclosure of Manual Alteration, section 3.4) analyzes workarounds held as personal habit within the answers on the three counted methods.

“

it hit the corporate card feed and pulled transactions for a cardholder that was not in the batch. nobody told it to. that happened twice in january. on cost, it will loop on the same receipt eight or nine times before it gives up, and there is no counter anywhere that shows that. i found it by watching the runtime go up. logs show it finished. thats all they show”

INDIVIDUAL CONTRIBUTOR

BPO, Corporate and back-office functions

2 Remedies Proposed by Respondents

The proposed remedies addressed operational responsibility, shared evidence, and timely authority. They included named people able to stop an agent, one incident record, written limits on autonomous action, and ownership of input definitions. These proposals connect the paper's reported weaknesses with possible control responses. They are suggestions from respondents, not interventions whose effectiveness the study evaluated. Remedies appeared in 225 answers from 203 respondents on 118 engagements, including 95 answers on canceled engagements.

Remedies were proposed by 27.56% (78 of 283) of respondents with a score of 3 or below, 21.02% (111 of 528) above 3 to 7, and 35.90% (14 of 39) above 7, the highest share of any score band. Portfolio directors proposed them at 41.94% (13 of 31), engagement managers at 26.00% (39 of 150), and individual contributors at 22.64% (151 of 667), and no segment family differed after Holm adjustment.

Of the 225 answers, 142 were judged negative on at least one governance indicator, and the indicators named the control the remedy would restore. Co-management indicators appeared in 56 of the 142 and financial institution readiness indicators in 54, against 20 for AI technology, 18 for the third-party ecosystem, and 6 for vendor capability. The indicators named most often were the financial institution's speed of response to issues raised (19), its provisioning of access and identities for agents (18), the effectiveness of its management of its side (16), the maturity of its data governance (12), and the placement of human review points (12). The remedies asked for one person on each side with authority to stop an agent, one shared list of incidents in place of two, a written statement of what an agent may do without a person, and one owner on the financial institution's side for the definitions the agents read.

> *“How I feel is nobody can press stop. What would help is simple, one person on each side with authority to pause the run, and the check should sit before the payment file goes out not after”*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> BPO, Corporate and back-office functions

> *“two separate queues is the real issue. one shared incident list would fix most of it, at present their soc raise their own ticket and we raise ours and nobody reconciles them.”*
>
> **ENGAGEMENT MANAGER**
>
> ITO, Technology engineering and operations

> *Practically, one page is needed: which program parameters an agent may set without a human, agreed by client treasury and by our operations together. Until that is signed, our reviewers approve everything on trust and the review point is decoration"*
>
> **ENGAGEMENT MANAGER**
>
> BPO, Payments, cards and disputes

Taken together, the findings place the operation and visibility of controls at the center of the governance problem. Lower scores were associated with canceled engagements and substantial manual correction, while specific reported weaknesses also appeared among more favorable composite assessments. Co-management showed consistent conditional associations, but category composition and model sensitivity limit claims of statistical uniqueness. The evidence supports examining who can decide, stop, review, and reconstruct agent work across organizational boundaries. It does not establish causal effects or certify effective governance at a numeric threshold. The respondents' remedies address those operating responsibilities and remain proposals to be evaluated in practice.

Part 5 (Leading Dynamics of Manual Alteration, section 10) analyzes the remedies proposed by the delivery team within the answers on manual alteration.

> *I pulled the field mappings myself, the same policy field is named three ways across their systems and nothing lines up. What would help is a single owner over there for the definitions... right now we ask four people and get four answers."*
>
> **INDIVIDUAL CONTRIBUTOR**
>
> ITO, Client servicing and communications

> *Incident last cycle: we pull month of call recordings and half have no adviser ID or call reason on them, so my team spends two nights re-tagging by hand. People are willing on client side, thats not issue. But I am the one explaining to our own leadership why throughput is short, and my QA leads are burning out on work nobody counts."*
>
> **ENGAGEMENT MANAGER**
>
> CXM, Regulatory, audit, tax and legal

About the Author

Phil Hatch

Outsourcing and workforce development

Akholi author page

LinkedIn

ORCID 0000-0003-1709-8860

SSRN research paper

Phil Hatch is an outsourcing and workforce-development executive with more than 30 years of experience. He previously led the global outsourcing business of ManpowerGroup/Experis. His work includes service delivery, workforce strategy, and organizational development across industries and countries.

Hatch has also advised governments on labor and economic development, including workforce strategies intended to support investment and employment.

Following Russia's 2022 invasion of Ukraine, Hatch was asked by the U.S. Department of State and USAID to advise Ukraine on workforce and economic development. His work has addressed technology, outsourcing, skills development, and international investment in support of economic recovery.

Through Akholi, Hatch advises organizations and governments on outsourcing and the implications of AI for work and service delivery.

“

The 6 on the shared risk view is narrower than it looks. We keep our risk log, they keep theirs, and the two only meet at the monthly forum, so a latency exposure we raised in week one sat on our side alone until the next cycle came round. The rest of the joint model holds. Who may restart what without asking us was settled early and has not caused an argument since.”

ENGAGEMENT MANAGER

ITO, Technology engineering and operations

About Akholi

Akholi, established in 2003, is a research and advisory firm focused on outsourcing, service delivery, and global capability centers. Its work includes the 2026 study of 150 agentic AI outsourcing engagements described in this report.

The study examines relationships among financial institution readiness, co-management, vendor capability, AI technology, the third-party ecosystem, and reported engagement outcomes. Its findings are observational and do not establish that any one category is the sole or primary cause of those outcomes.

Akholi applies this research through advisory and educational work with institutional leaders. Recommendations derived from the study require assessment against the circumstances and evidence of each engagement.

Akholi

PART 3 / 2026 RESEARCH STUDY

Governance, Oversight, and Control

Part 3 examines how governance, oversight, and control operate across organizational readiness, joint management, technical controls, third-party dependencies, obligations, reporting, and manual alteration.

Phil Hatch | Akholi

Study ISBN 979-8-234-26882-2
SSRN DOI: https://doi.org/10.2139/ssrn.7164559
Figshare DOI: https://doi.org/10.6084/m9.figshare.34036890

Full study and supporting data

akholi.com/outsourcing-2026-agentic-ai-tier-1-banks

Version 1.8.2 | November 1, 2026

The evidence reflects respondents' reported experience. It is not an independent finding of regulatory breach.