INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EVOLUTIONARY RESEARCH

From Checklists to Exposure Estimates: A Governance-Anchored Framework for AI-Enabled Third-Party Risk Quantification in Multi-Entity Supply Chain Environments

Olakunle Akintunde Akinbowale
Faculty of Management Sciences, University of Benin, Nigeria

* Corresponding Author: **Olakunle Akintunde Akinbowale**

Article Info

P-ISSN: 3051-3502
E-ISSN: 3051-3510
Volume: 03
Issue: 01
January - June 2022
Received: 08-12-2021
Accepted: 06-01-2022
Published: 04-02-2022
Page No: 123-138

Abstract
Third-party dependencies have become a structural feature of contemporary supply chain operations, creating exposures that extend beyond the boundaries of any single organization. Prevailing assessment practice remains anchored in periodic, questionnaire-based methods that produce ordinal risk ratings. Such ratings cannot express uncertainty, cannot be aggregated across organizational entities, and decay between review cycles. Artificial intelligence offers analytical capabilities that address these constraints, yet the literature reviewed here treats those capabilities largely in isolation from the governance structures that determine whether their outputs acquire organizational force. This paper develops a governance-anchored framework for AI-enabled third-party risk quantification through design science research, positioning the contribution as an exaptation of quantification approaches established in credit risk to a domain characterized by sparse outcome data. The framework comprises four layers: data ingestion and signal extraction, quantification logic, governance architecture, and aggregation and escalation. Governance controls are specified within it as constitutive components rather than as subsequent additions, and its quantification layer is scoped to relative exposure ordering with explicit uncertainty rather than to calibrated absolute failure probabilities, a scoping decision that follows from the rare-event data conditions characteristic of vendor populations. Evaluation proceeds against external referents comprising named regulatory expectations and documented failure cases, following an established design science evaluation framework. Of six design requirements, four are assessed as adequately addressed, one as partially addressed, and one as conditionally addressed pending specification beyond the framework's present scope. The study contributes to supply chain risk management, enterprise risk governance, and applied artificial intelligence scholarship. As a conceptual artifact evaluated analytically rather than empirically, its propositions await implementation testing.

DOI: https://doi.org/10.54660/IJMER.2022.3.1.123-138

Keywords: Third-party risk management, AI-enabled risk quantification, supply chain governance, enterprise risk management, design science research

1. Introduction

Over the past two decades, the fragmentation and globalization of supply chains has reshaped the risk landscape facing industrial and service organizations (Christopher and Peck, 2004; Chopra and Sodhi, 2004) [6, 5]. Where firms once managed production, logistics, and support functions through vertically integrated structures, the prevailing model relies on networks of third-party providers, each contributing specialized capability while introducing exposures beyond the contracting organization's direct control (Bode and Wagner, 2015) [3].

Reviews of the supply chain risk literature document both the growth of this dependence and the analytical attention it has attracted (Ho *et al*., 2015; Fan and Stevenson, 2018) [21, 18]. Supervisory bodies have responded with instruments and expectations directed at ensuring that organizations identify, assess, and monitor the risks arising from external relationships (OCC, 2013; NIST, 2015).

The assessment practices through which organizations discharge these expectations have evolved more slowly than the exposures they address. Periodic questionnaires, self-reported attestations, and ordinal risk ratings remain the dominant instruments through which vendor risk is assessed and communicated. These instruments are administratively tractable, and their persistence is not accidental: they align with established audit rhythms and produce documentation that satisfies external review (Power, 1997, 2007). They also carry limitations that compound in organizations operating across multiple legal entities, geographies, or regulatory jurisdictions. A vendor rated medium risk by one entity may carry materially different implications for another, and the ordinal format supplies no mechanism for distinguishing the underlying exposures or combining them into a group-level view.

Artificial intelligence offers analytical capabilities that bear directly on these limitations. Textual analysis of corporate disclosures has demonstrated that risk-relevant signals can be extracted systematically from unstructured documents at scale (Loughran and McDonald, 2011, 2016). Machine learning approaches to credit risk have shown that quantitative exposure estimation is achievable in domains with sufficient outcome data (Khandani *et al*., 2010; Leo *et al*., 2019). Whether these capabilities transfer to third-party risk assessment is not, however, a purely technical question. A data condition and a governance condition both stand in the way, and the literature reviewed here addresses neither adequately.

The data condition is the more fundamental. Credit risk quantification succeeded on a base of millions of observed defaults, whereas vendor failure is comparatively rare, and organizations observe too few instances within their own portfolios to calibrate failure probabilities directly. Any proposal to quantify third-party exposure must state a position on this constraint rather than proceeding as though it were absent.

The governance condition operates in both directions. A quantification capability operating outside a governance structure produces estimates with no defined pathway to organizational action, while a governance structure without quantification capability perpetuates dependence on the ordinal methods whose limitations prompted the inquiry. Existing scholarship on AI in risk assessment concentrates on the performance characteristics of models rather than the accountability arrangements within which model outputs acquire organizational force (Rudin, 2019; Raji *et al*., 2020). Drawing on these observations, the study addresses the following research question: how can AI-enabled capabilities be integrated into governance-anchored frameworks to transform third-party risk assessment from ordinal classification to probabilistic exposure quantification in multi-entity supply chain environments?

The study adopts design science research (Hevner *et al*., 2004; Peffers *et al*., 2007) to develop a framework connecting supply chain risk management, enterprise risk governance, and applied artificial intelligence. Following the knowledge contribution taxonomy of Gregor and Hevner (2013), the contribution is positioned as an exaptation: the extension of quantification approaches developed in data-rich financial risk domains to a domain where outcome data is sparse and governance requirements are stringent. Exaptation requires demonstrating that the transferred approach must be modified to suit the new setting, and the modifications proposed here concern both the scope of the quantification claim and the governance apparatus surrounding it.

The contribution is threefold. The framework integrates AI capabilities within governance architecture as constitutive rather than supplementary components. It proposes a quantification approach scoped to what sparse-data conditions permit, producing relative exposure ordering with explicit uncertainty rather than calibrated absolute probabilities. And it specifies the governance design requirements, including escalation logic, override protocols, and assurance of the AI components themselves, that condition whether AI-enabled quantification functions as an accountable organizational capability.

Section 2 reviews the theoretical background. Section 3 describes the methodology, including the literature protocol and evaluation design. Section 4 presents the framework. Section 5 demonstrates its application through two documented failure cases. Sections 6 and 7 discuss and conclude.

2. Theoretical Background

2.1. Third-party risk in contemporary supply chains

Supply chain risk has attracted sustained scholarly inquiry since the early 2000s, driven by recognition that the benefits of specialization and global sourcing carry exposures that organizations underestimate (Chopra and Sodhi, 2004; Zsidisin, 2003) [5]. Zsidisin (2003, p. 222) defined supply risk as "the probability of an incident associated with inbound supply from individual supplier failures or the supply market occurring, in which its outcomes result in the inability of the purchasing firm to meet customer demand or cause threats to customer life and safety." Subsequent research identified the structural drivers of such risk and the organizational capabilities required to manage it, with Craighead *et al*. (2007) establishing that disruption severity depends on supply chain design characteristics including density, complexity, and node criticality, and on the mitigation capabilities of recovery and warning.

Structural complexity has intensified these concerns. Bode and Wagner (2015) [3] demonstrated that upstream supply chain complexity, operationalized through horizontal, vertical, and spatial dimensions, is positively associated with disruption frequency. The extension of outsourcing beyond manufacturing and logistics to encompass information technology, financial processing, and compliance functions has created dependencies that cut across industry boundaries. The COVID-19 pandemic exposed the fragility of supply chains optimized for efficiency at the expense of resilience, with disruptions originating in geographically concentrated supplier bases cascading through dependent industries in patterns existing instruments had not anticipated (Ivanov, 2020).

Within this literature, third-party risk denotes the exposure arising from an organization's direct contractual relationships with external providers of goods, services, or critical functions. Financial services supervisors have formalized expectations in this area, establishing that outsourcing a

critical function does not transfer the associated regulatory responsibility and requiring ongoing monitoring rather than assessment at fixed intervals alone (OCC, 2013). Comparable expectations have emerged for federal information systems, where dependencies on external technology providers create exposures that propagate across organizational boundaries (NIST, 2015).

Assessment practice has not kept pace. The scale and interconnectedness of contemporary vendor portfolios have outpaced the capacity of traditional methods, producing a widening distance between the complexity of the risk landscape and the precision of the instruments used to navigate it.

2.2. Prevailing assessment approaches and their limitations

The dominant approach combines periodic questionnaires, self-reported attestations, and ordinal risk ratings. Vendors are evaluated at onboarding and at prescribed intervals through structured instruments covering security controls, financial stability, operational continuity, and regulatory compliance. Responses are scored, weighted, and translated into ordinal classifications, commonly a three- or five-tier scale.

Table 1 states the four limitations that prove consequential for multi-entity environments, alongside their consequences and supporting sources.

Table 1: Structural limitations of prevailing third-party risk assessment approaches

Limitation	Characteristic	Consequence for Multi-Entity Environments	Supporting Sources
L1: Periodicity	Assessment at fixed intervals produces point-in-time evidence that ages between cycles, while supervisory expectations require ongoing monitoring	Exposure windows widen as operating volatility increases; assessment intervals do not adjust to changing vendor conditions; entities on different review calendars hold inconsistent views of the same vendor	OCC (2013); Power (2007)
L2: Information asymmetry	Reliance on vendor self-reported data, where the assessed party controls disclosure and holds superior information about its own condition	Certification scope is vendor-defined and may not cover the services generating a given entity's exposure; verification cost rises with portfolio size	Eisenhardt (1989); Power (1997)
L3: Ordinal non-aggregability	Ordinal ratings treat continuous underlying quantities as discrete classes, producing categories that cannot be combined arithmetically and that can invert the ranking of the quantities they represent	Identical category labels applied under different instruments, weightings, and risk appetites produce outputs sharing vocabulary but carrying no common meaning; group-level aggregation is arithmetically undefined	Cox (2008); Kaplan and Garrick (1981)
L4: Individual-vendor framing	Each relationship assessed as an independent unit, with dependence between vendors unmodelled	Concentration, shared sub-tier dependencies, and correlated failure modes remain invisible at portfolio and group level; an entity-diversified portfolio may carry group-level concentration	Bode and Wagner (2015) [3]; Christopher and Peck (2004) [6]

Ordinal non-aggregability (L3) and individual-vendor framing (L4) carry disproportionate weight in the argument that follows, and warrant elaboration.

The ordinal aggregation problem is not merely a matter of imprecision. Cox (2008) demonstrated that risk matrices constructed from ordinal categories can assign identical ratings to quantitatively very different risks, and can under specified conditions rank a lower quantitative risk above a higher one. The underlying difficulty is that ordinal categories discard the information required for coherent combination. Kaplan and Garrick's (1981) formulation of risk as a set of scenario, likelihood, and consequence triplets makes explicit what ordinal ratings suppress: the likelihood component is a continuous quantity, and collapsing it into a category forecloses the arithmetic on which aggregation depends. Figure 1 illustrates the information loss schematically.

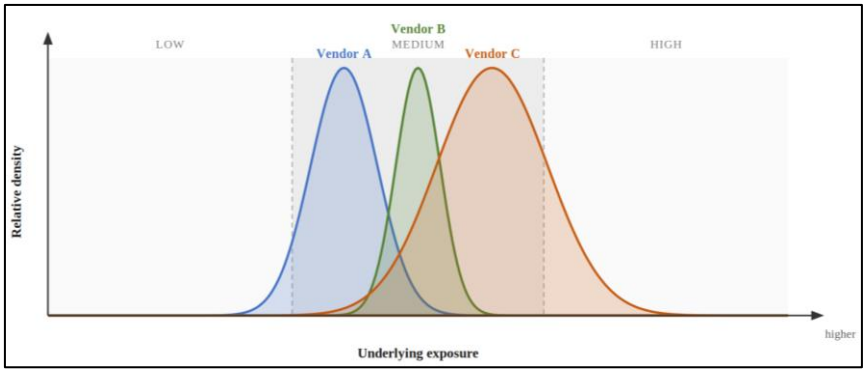

Fig 1: Schematic illustration of the information discarded by ordinal classification. The distributions are illustrative rather than empirical, representing the property established analytically by Cox (2008)

The consequences are visible in the illustration. Vendors whose central exposure differs materially receive the same label, so the classification does not discriminate where discrimination matters. Vendors whose uncertainty differs materially receive the same label, so a well-characterized exposure is indistinguishable from a poorly-characterized one. And vendors sitting close to a category boundary are indistinguishable from those sitting at its centre, so small

movements in underlying exposure produce either no change in classification or a discontinuous jump. In a single-entity setting these costs may be tolerable. Across entities, where the objective is a consolidated view, they are disabling.

The individual-vendor framing problem (L4) is the point at which third-party risk assessment departs most sharply from the supply chain risk literature's own findings. That literature has established for two decades that disruption propagates through network structure rather than through isolated nodes (Christopher and Peck, 2004; Craighead *et al*., 2007) [6]. Assessment practice nonetheless remains organized around the individual vendor as unit of analysis. An organization may rate each critical vendor as low or medium risk while carrying substantial aggregate exposure through structural dependencies that individual assessments do not surface.

These limitations are not primarily technological, which is why introducing more capable analytical instruments does not by itself resolve them. They reflect the governance architecture within which assessment takes place, a point developed in Section 2.3.

2.3. Enterprise risk governance and the conditions for quantification

Enterprise risk management frameworks advanced by the Committee of Sponsoring Organizations (COSO, 2017) and the International Organization for Standardization (ISO 31000, 2018) provide principles for integrating risk management across organizational functions and decision processes. Empirical scholarship has been more equivocal about what such frameworks achieve. Bromiley *et al*. (2015) observed in their review that ERM research has generated limited evidence connecting framework adoption to improved outcomes, and Arena *et al*. (2010) traced how ERM systems evolve within organizations in ways that reflect internal political dynamics as much as technical design.

Power (1997, 2007) supplies the sharpest version of this scepticism, and the version this paper must answer. His argument across *The Audit Society* and *Organized Uncertainty* is that assurance and risk management apparatus proliferates because it satisfies demands for accountability and auditability, not because it improves outcomes. Formal systems acquire ceremonial character: they generate documentation, withstand inspection, and reassure external parties, while the substantive questions they nominally address go unexamined.

This argument bears directly on the present study, and the tension must be stated rather than left implicit. A paper proposing an additional layer of risk apparatus, complete with accountability matrices, escalation protocols, and assurance regimes, is proposing precisely the kind of elaboration Power warns against. The framework developed here could plausibly become another ceremonial system: continuously ingesting data, producing probabilistic outputs, generating audit trails, and changing nothing.

The framework answers this objection through features whose adequacy remains a fair question for evaluation rather than an assumption. Its assurance protocol in Layer 3 subjects the analytical layers to periodic review against outcomes, creating a feedback path that ceremonial systems characteristically lack. Override decisions are recorded with rationale and monitored for pattern, so that systematic divergence between model output and human decision becomes visible rather than absorbed. Escalation triggers are specified in advance and tied to defined thresholds, which constrains the discretion through which inconvenient signals are ordinarily neutralized. And the governance layer is itself subject to assurance originating outside the framework, so that the apparatus cannot certify its own substantive operation, which is the mechanism by which ceremonial systems ordinarily sustain themselves. Whether this suffices is addressed in Section 6.1; the objection is not disposed of by naming them.

The distinction between qualitative and quantitative approaches has been persistent within this literature. Kaplan and Mikes (2012) distinguished preventable risks, managed through rules and compliance, from strategy risks accepted in pursuit of return and external risks arising beyond organizational control, arguing that each category requires a different management approach. Third-party exposures span all three categories while prevailing assessment treats them uniformly through a single ordinal instrument.

The paper uses the term governance-anchored in a specific sense that warrants definition. A risk quantification capability is governance-anchored when three conditions hold jointly: accountability for acting on its outputs is assigned to identified roles in advance of those outputs being produced; the thresholds at which outputs trigger escalation are specified in advance rather than negotiated after the fact; and the quantification instruments are themselves subject to defined assurance, with authority to challenge their outputs vested somewhere other than the function that operates them. A capability satisfying none of these conditions may be analytically sophisticated while remaining organizationally inert. The definition is deliberately demanding, and it is the standard against which Layer 3 is evaluated.

2.4. The multi-entity condition

Organizations operating through multiple legal entities face third-party risk conditions that single-entity treatments do not capture, and this dimension has received limited attention in the assessment literature reviewed here.

Risk appetite in the multi-entity case is set at group level but applied at entity level, and the translation between them is imperfect: an exposure tolerable within a large entity's appetite may breach a smaller entity's, while the group view depends on an aggregation that ordinal ratings cannot support. Vendor relationships are frequently shared across entities without a defined owner, so that a provider material to three entities may be assessed by each independently, with none holding the consolidated view and none formally accountable for the relationship as a whole. Concentration accumulates invisibly, because an entity-level portfolio that appears diversified may combine with sibling entities' portfolios into a group-level concentration that no entity-level assessment surfaces. And supervisory reporting obligations frequently attach at group level while the assessment evidence supporting them is generated at entity level under divergent instruments and calibrations.

These features convert the limitations of Section 2.2 from inefficiencies into structural failures. Periodicity (L1) becomes inconsistency when entities operate different review calendars. Ordinal non-aggregability (L3) becomes an inability to produce the group view that supervisors expect. Individual-vendor framing (L4) becomes blindness to concentration that exists only at group level and is therefore invisible everywhere it is assessed. The framework

developed in Section 4 treats the multi-entity condition as its primary design context rather than as an extension of the single-entity case.

2.5. AI capabilities in risk assessment, and the sparse-data constraint

Applications of machine learning to risk assessment have expanded across credit risk (Khandani *et al*., 2010), fraud detection (Ngai *et al*., 2011), and financial risk management (Leo *et al*., 2019). Textual analysis has developed in parallel, with Loughran and McDonald (2011) establishing that domain-specific lexicons materially outperform general-purpose ones in extracting sentiment and risk signals from corporate filings, and their subsequent survey (Loughran and McDonald, 2016) documenting the maturation of textual methods across accounting and finance. Advances in language representation extended the technical capability available for such tasks (Devlin *et al*., 2019).

For third-party risk assessment, these capabilities bear on three of the four limitations identified above. Textual analysis of vendor disclosures, regulatory filings, contractual terms, and audit opinions can extract structured signals from documents that presently require manual review, addressing periodicity (L1) by permitting processing as disclosures appear rather than at scheduled intervals. Ingestion of externally sourced data, including regulatory actions, litigation records, credit rating changes, and market indicators, supplies information the assessed party does not control, addressing information asymmetry (L2). Network-level analysis across a vendor portfolio addresses individual-vendor framing (L4) by making dependence structure an object of analysis.

The third limitation, ordinal non-aggregability (L3), is the one AI capability does not straightforwardly resolve, and the reason is a data condition that must be confronted directly.

Quantitative risk estimation in credit markets succeeded on an outcome base of substantial size: consumer default is frequent, labelled, and recorded over long periods, which permits models to be calibrated and validated against realized outcomes (Khandani *et al*., 2010). Vendor failure does not share these properties. Outright failures are rare within any single organization's portfolio, are heterogeneous in form, and are recorded inconsistently where they are recorded at all. An organization with two thousand vendors may observe a handful of material failures across a decade. That base cannot support calibration of absolute failure probabilities, and no volume of signal ingestion changes this, because the constraint lies in the outcome data rather than the predictor data.

This constraint has three consequences for the framework's design, and stating them is necessary for the quantification claim to be credible.

The claim must be scoped to what the data permits. The framework does not propose calibrated absolute failure probabilities. It proposes relative exposure ordering across a vendor portfolio, expressed with explicit uncertainty. Relative ordering requires only that the quantification be monotonic in exposure, a substantially weaker requirement than calibration, and it is sufficient for the practical purposes of prioritization, threshold-based escalation, and concentration analysis.

Outcome definition must be broadened beyond failure. Sparse-outcome problems are commonly addressed by substituting more frequent proxy events that lie on the causal path to the rare event of interest. In this domain, service level breaches, contract disputes, remediation demands, unplanned substitution, adverse regulatory findings, and material deterioration in financial condition occur with sufficient frequency to support estimation, and they carry information about the rare outcome. The framework's estimation target is therefore a composite adverse-event construct rather than failure alone.

Estimation must combine judgment with observation. Where outcome data is sparse, structured expert elicitation supplies prior distributions that sparse observations update, a methodology developed extensively for exactly these conditions (O'Hagan *et al*., 2006). Aven's (2016) treatment of the foundations of risk assessment supports the associated interpretive position: probabilities in such settings express knowledge-conditional judgment rather than observed frequency, and this should be stated rather than obscured. Hierarchical estimation, pooling information across vendors sharing sector, size, or service characteristics, extends the effective evidence base for any individual vendor.

The governance implications of automated risk classification constitute a separate concern. Supervisory guidance on model risk management, developed for the financial sector, establishes expectations for validation, documentation, and independent challenge of models used in consequential decisions (Federal Reserve and OCC, 2011). That guidance predates current AI capabilities but its structure remains applicable. Parallel work in algorithmic accountability has proposed internal audit methodologies for AI systems (Raji *et al*., 2020), documentation standards for model reporting (Mitchell *et al*., 2019), and cautions against deploying opaque models in high-stakes decisions where interpretable alternatives exist (Rudin, 2019). Emerging AI governance frameworks articulate transparency, accountability, and fairness as deployment conditions (Floridi *et al*., 2018; Jobin *et al*., 2019). This body of work supplies the governance requirements the framework's assurance layer must satisfy, and its concentration in general-purpose AI governance rather than in third-party risk specifically is what the framework attempts to bridge.

2.6. Research positioning

Within the literature reviewed, three developments converge to define the research space. Structural dependence on third parties has created exposures whose complexity exceeds the capacity of prevailing assessment methods (Chopra and Sodhi, 2004; Bode and Wagner, 2015; Ivanov, 2020) [5, 3]. Enterprise risk governance frameworks establish principles for integrating risk assessment into decision-making without prescribing the mechanisms through which qualitative inputs become quantitative outputs in third-party contexts (COSO, 2017; ISO 31000, 2018; Bromiley *et al*., 2015). AI capabilities supply analytical means for that conversion, but their treatment concentrates on model performance rather than on the governance arrangements that determine organizational effect (Rudin, 2019; Raji *et al*., 2020).

The intersection is underdeveloped in the literature reviewed here. Supply chain risk scholarship documents the drivers and consequences of third-party exposure with limited attention to the assessment methods through which exposure is quantified at organizational level (Ho *et al*., 2015; Fan and Stevenson, 2018) [18]. Enterprise risk management literature

articulates the qualitative-quantitative distinction without examining its application to third-party relationships specifically (Bromiley *et al*., 2015). Machine learning scholarship demonstrates technical feasibility without situating capability within the governance structures determining organizational legitimacy (Leo *et al*., 2019).

This assessment is bounded in two respects stated in Section 3.3. The knowledge base was assembled purposively rather than through systematic search, so the assessment characterizes the literature the study engages rather than establishing the absence of contrary work. And the source boundary closes at the end of 2021, so the assessment cannot speak to the field's present state. The positioning claim should be read within both limits, and Section 7 addresses their implications.

3. Methodology

3.1. Research approach and its justification

Design science research supplies the methodological foundation, following Hevner *et al*. (2004) and the process model of Peffers *et al*. (2007). It concerns the creation and evaluation of artifacts addressing identified classes of organizational problems, and is distinguished from explanatory research by its prescriptive orientation.

The choice follows from the research question. The study does not seek to explain why prevailing assessment methods persist, which would call for behavioural or institutional analysis. It seeks to develop an artifact addressing their structural limitations. Table 2 maps the study against the seven guidelines of Hevner *et al*. (2004), a mapping included because design science submissions are frequently criticized for invoking the methodology without applying it.

Table 2: Study design against Hevner *et al*. (2004) design science guidelines

Guideline	Requirement	Treatment in this study
1. Design as an artifact	Produce a viable artifact	A four-layer framework specified in Section 4 with defined inputs, outputs, controls, and interfaces
2. Problem relevance	Address an important organizational problem	Third-party risk assessment inadequacy, evidenced through literature and supervisory expectation (Sections 2.1 to 2.4)
3. Design evaluation	Demonstrate utility, quality, and efficacy through well-executed evaluation	Analytical and descriptive evaluation against external referents following Venable *et al*. (2016), reported in Section 6.1
4. Research contributions	Provide clear, verifiable contributions	Exaptation contribution positioned per Gregor and Hevner (2013), stated in Section 6.2
5. Research rigour	Apply rigorous methods in construction and evaluation	Knowledge base selection stated with basis for each source category (Section 3.3); design theory components per Gregor and Jones (2007); evaluation design per Venable *et al*. (2016)
6. Design as a search process	Use available means to reach desired ends within problem constraints	Design requirements derived from identified limitations; sparse-data constraint treated as a binding condition shaping the solution scope (Section 2.5)
7. Communication of research	Present to both technical and managerial audiences	Framework specified at implementation-relevant granularity (Section 4); adoption conditions addressed (Section 6.4)

3.2. Design science protocol and requirements derivation

The study follows the six activities of Peffers *et al*. (2007). Problem identification and motivation proceed through the literature review reported in Section 2 and the supervisory instruments identified there. Definition of objectives translates identified limitations into design requirements. Design and development produce the framework in Section 4. Demonstration applies the framework to documented cases in Section 5. Evaluation assesses the artifact against external referents in Section 6.1. Communication comprises the present article.

Table 3 states the design requirements and their derivation. Requirements DR1 through DR4 correspond to the limitations of Table 1. DR5 derives from the governance conditions of Section 2.3, and DR6 from the sparse-data constraint of Section 2.5.

Table 3: Design requirements, derivation, and framework allocation.

Requirement	Origin	Statement	Framework layer
DR1: Continuity	L1	Support assessment that updates as evidence emerges rather than at fixed intervals, consistent with ongoing monitoring expectations	Layer 1
DR2: Independence	L2	Incorporate evidence sources the assessed party does not control	Layer 1
DR3: Quantification	L3	Produce exposure estimates supporting arithmetic combination and comparison, with uncertainty represented explicitly	Layer 2
DR4: Portfolio and group awareness	L4, Section 2.4	Represent dependence between vendors and support consolidation across organizational entities under divergent local calibration	Layer 4
DR5: Governance anchoring	Section 2.3	Satisfy the three conditions of governance anchoring: advance accountability assignment, advance threshold specification, and independent assurance of the quantification instruments	Layer 3
DR6: Data-condition realism	Section 2.5	Scope quantification claims to what sparse outcome data supports, and specify the estimation approach accordingly	Layer 2

3.3. Knowledge base and source selection

Design science research draws on an existing knowledge base of theories, methods, and established artifacts, and the rigour of a design study depends on the appropriateness of what it draws upon rather than on exhaustive coverage of a field (Hevner *et al*., 2004; Hevner, 2007). The knowledge base

here was assembled purposively rather than through systematic search, and the basis of selection is stated so that the foundation of each element of the argument can be traced and contested.

Sources fall into four categories, each serving a distinct function in the design.

Theoretical foundations establish the constructs on which the problem statement rests. These were selected as the works that define the constructs in question rather than as a representative sample of the field: Zsidisin (2003) for supply risk, Christopher and Peck (2004) [6] and Craighead *et al.* (2007) for disruption and network structure, Kaplan and Garrick (1981) for the formal structure of risk statements, and Power (1997, 2007) for the account of risk apparatus as organizational practice. Review articles were used to characterize the state of each field rather than to substitute for engagement with primary work (Ho *et al*., 2015; Fan and Stevenson, 2018; Bromiley *et al*., 2015) [18].

Methodological standards govern the techniques the artifact employs and supply the criteria against which its methodological claims are assessed. Selection was determined by the artifact's design rather than by field coverage: because the quantification layer proposes elicited priors under sparse outcome data, O'Hagan *et al*. (2006) applies; because it aggregates across dependent units, Embrechts *et al*. (2002) applies; because it displaces ordinal rating, Cox (2008) applies; because it deploys models in consequential decisions, the algorithmic accountability literature applies (Raji *et al*., 2020; Mitchell *et al*., 2019; Rudin, 2019).

Supervisory instruments and standards define requirements the artifact must satisfy and supply the external referents used in evaluation. These were retrieved directly from issuing authorities and comprise OCC Bulletin 2013-29, NIST SP 800-161, ISO 31000:2018, COSO ERM (2017), and the interagency supervisory guidance on model risk management (Federal Reserve and OCC, 2011). Selection was determined by applicability to third-party risk oversight and to the governance of models used in risk decisions.

Documented failure cases supply the second evaluation referent. Selection required a public investigative record produced by a body independent of the parties involved and independent of this study, and contrasting failure modes across the cases chosen. The two cases examined in Section 5 satisfy both conditions, resting on national audit and parliamentary investigation reports and on a government cybersecurity advisory.

Sources were bounded to publications available to 31 December 2021. The boundary is a scope condition adopted to characterize the state of knowledge preceding the widespread availability of general-purpose generative language capability, a development that altered the technical and supervisory landscape sufficiently to warrant separate treatment rather than incorporation. Section 7 states its consequences.

This approach carries a cost that warrants acknowledgment rather than defence. The knowledge base is not the product of systematic search, and the study accordingly makes no claim to exhaustive coverage of any of the four literatures. Claims about the state of a field are correspondingly bounded, and Section 2.6 states the resulting limit on the study's positioning claim. What the approach does support is traceability: each element of the argument is anchored to an identified source selected for a stated reason, which permits a reader to assess whether the anchor bears the weight placed on it. In a design study, whose warrant rests on the coherence of the design logic and on performance against external referents rather than on the completeness of a literature census, this is the property that matters (Hevner, 2007).

3.4. Evidence base

Peer-reviewed literature and scholarly monographs, identified as described above, supply the study's theoretical evidence. Supervisory instruments and international standards supply its normative evidence, specifically OCC Bulletin 2013-29, NIST SP 800-161, ISO 31000:2018, COSO ERM (2017), and the interagency supervisory guidance on model risk management (Federal Reserve and OCC, 2011). Publicly documented third-party failure cases, drawn from national audit and parliamentary investigation reports and from government cybersecurity advisories, supply its empirical evidence and are used for demonstration in Section 5.

No primary data were collected. The framework's warrant rests on the coherence of its design logic relative to the identified problem and requirements, and on its performance against the external referents specified in Section 3.5, rather than on empirical generalization.

3.5. Evaluation design

Evaluation follows the framework of Venable *et al*. (2016), which distinguishes evaluation by purpose (formative or summative) and by setting (artificial or naturalistic). The evaluation reported here is formative and artificial: it assesses a newly constructed artifact analytically rather than through deployment. This positioning is appropriate for an initial design cycle and is stated explicitly because summative naturalistic evaluation, requiring implementation, is beyond the study's scope.

Assessment against requirements the artifact's designer authored would be circular and would establish nothing. The evaluation therefore applies external referents, assessing each design requirement against whichever of the three below are relevant to it.

Referent 1: Supervisory expectation. Does the framework layer satisfy requirements stated in instruments the study did not author? Specifically, the ongoing monitoring and risk-based oversight expectations of OCC Bulletin 2013-29; the supply chain risk control expectations of NIST SP 800-161; the risk process requirements of ISO 31000:2018; and the model validation, documentation, and independent challenge expectations of the interagency model risk guidance.

Referent 2: Documented failure evidence. Does the framework layer address mechanisms identified in the independent investigative record of the cases examined in Section 5? The reports relied upon were produced by national audit and parliamentary bodies and by government cybersecurity authorities, none with any relation to this study.

Referent 3: Established methodological standards. Where the framework makes methodological claims, do those claims conform to standards established in the relevant literature? Applied to the quantification approach against the risk quantification literature, and to the assurance protocol against the algorithmic accountability literature.

The categories applied are differentiated rather than uniform. A requirement is **adequately addressed** where the framework specifies a mechanism that satisfies the relevant referents. It is **partially addressed** where a mechanism is

specified but leaves identified conditions unmet. It is **conditionally addressed** where the framework specifies what must be determined without determining it, so that satisfaction depends on decisions outside the artifact's present scope. Section 6.1 reports the results.

3.6 Trustworthiness

Problem identification rests on independently verifiable literature and named supervisory instruments rather than assumed conditions. The basis for selecting each category of source is stated in Section 3.3, so that a reader can assess whether a given anchor bears the weight the argument places on it, though purposive selection means the study cannot claim that contrary evidence was sought and not found. Requirements are derived transparently from identified limitations, and Tables 1 and 3 make the derivation auditable. The framework is specified at a granularity permitting independent assessment. Evaluation referents are external to the study, which is the principal safeguard against the circularity to which single-designer evaluation is otherwise exposed. The design was conducted by a single researcher without inter-researcher validation, a limitation stated in Section 7 and partially offset by the traceability of the design chain, which permits others to assess each step independently.

4. Framework Development

The framework comprises four layers addressing the six design requirements of Table 3. Layers are sequentially dependent, each layer's outputs serving as the next layer's inputs, with an assurance path running from Layer 3 to the three analytical layers. Following Gregor and Jones (2007), the specification states each layer's purpose, constructs, principles of form and function, and the conditions on which its operation depends. Figure 2 presents the architecture and Table 4 the layer specification.

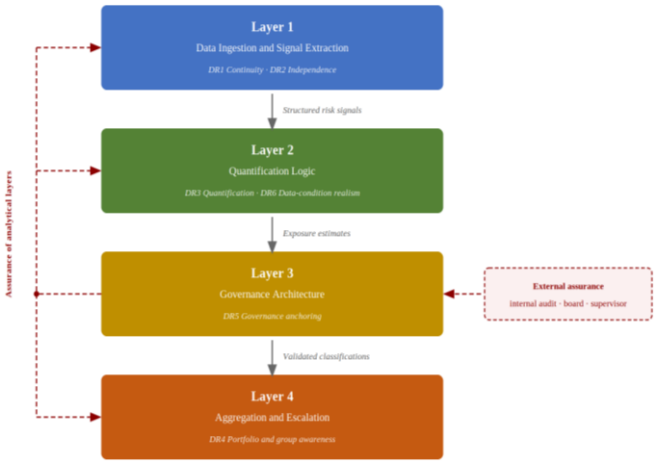

Fig 2: Architecture of the governance-anchored framework, showing four layers with sequential data flows, inputs and outputs, AI roles, the assurance path from Layer 3 to the analytical layers, and the external assurance on which Layer 3 itself depends

Table 4: Framework layer specification.

Layer	Inputs	Outputs	AI role (functional)	Governance controls
Layer 1: Data ingestion and signal extraction	Vendor disclosures, regulatory filings, financial data, litigation and enforcement records, market indicators, certification reports, contractual terms	Structured risk signals classified by category, with provenance and recency metadata	Textual extraction of risk-relevant content from unstructured documents; automated monitoring of external sources	Source admissibility rules; provenance recording; data quality and recency thresholds; minimum independent-source proportion
Layer 2: Quantification logic	Structured signals from Layer 1; historical adverse-event records; elicited priors; hierarchical pooling parameters	Relative exposure estimates with explicit uncertainty intervals, on a common scale across the portfolio	Signal-to-estimate mapping including non-additive signal interaction; anomaly detection against vendor and peer-group baselines	Elicitation protocol documentation; model validation against realized adverse events; weighting transparency; recalibration schedule; challenge authority independent of the operating function
Layer 3: Governance architecture	Exposure estimates from Layer 2; entity and group risk appetite; supervisory thresholds; escalation policy	Accountability-assigned classifications; escalation instructions; override records with rationale; assurance findings	Threshold monitoring and escalation triggering	Accountability assignment matrix fixed in advance; override rationale capture and pattern monitoring; scheduled assurance of Layers 1, 2, and 4 including bias, drift, interpretability, and dependence assumption review; external assurance of the layer itself
Layer 4: Aggregation and escalation	Classified estimates across entities; vendor dependency and sub-tier mapping; entity calibration parameters	Entity and group exposure views; concentration and correlated-exposure indicators; consolidated supervisory reporting inputs	Dependency inference across the vendor network; correlation and concentration detection	Cross-entity calibration protocol; aggregation methodology disclosure; dependence assumption documentation; group-level escalation thresholds

4.1. Layer 1: Data ingestion and signal extraction

Layer 1 addresses DR1 and DR2 by reconstituting the evidence base on which quantification depends. Under prevailing practice, the primary source is the vendor's own questionnaire response, supplemented by point-in-time certification. This produces a temporal constraint, since evidence refreshes only at prescribed intervals, and an independence constraint, since the principal informant is the assessed party (Eisenhardt, 1989).

The layer ingests two categories of evidence continuously. Vendor-originated disclosures comprise financial statements, regulatory submissions, contractual terms, breach notifications, and audit opinions. Textual analysis extracts structured signals from these as they appear rather than at the next scheduled review. The methodological basis is established: Loughran and McDonald (2011) demonstrated that domain-calibrated textual analysis of corporate filings extracts risk-relevant signal that general-purpose approaches miss, and the subsequent literature has extended these methods across financial and accounting applications (Loughran and McDonald, 2016). The framework's requirement is that extraction be domain-calibrated for vendor risk rather than transferred unmodified from financial sentiment applications, since the constructs of interest differ.

Externally sourced evidence comprises regulatory enforcement actions, litigation filings, credit rating changes, market indicators including short interest where the vendor is publicly traded, and public reporting. These sources supply information the vendor does not control, addressing DR2 directly. Their continuous ingestion means that material deterioration becomes visible within days rather than remaining latent until the next review cycle.

Governance controls address evidence integrity. Source admissibility rules specify which sources are sufficiently reliable to enter the pipeline. Provenance recording attaches source and timestamp to every signal, without which the assurance review in Layer 3 cannot trace an estimate to its evidence. Recency thresholds prevent stale evidence from carrying undue weight. A minimum independent-source proportion prevents the estimate from becoming dominated by vendor-controlled evidence, which would reinstate the asymmetry the layer exists to reduce.

4.2. Layer 2: Quantification logic

Layer 2 addresses DR3 and DR6, converting structured signals into exposure estimates. The layer's design is shaped throughout by the sparse-data constraint established in Section 2.5, and its claims are scoped accordingly.

Estimation target: The layer does not estimate calibrated absolute probabilities of vendor failure. It estimates relative exposure across the portfolio on a common continuous scale, with uncertainty represented explicitly as an interval rather than suppressed into a point value. The target construct is a composite adverse-event exposure encompassing service level breach, contract dispute, remediation demand, unplanned substitution, adverse regulatory finding, and material financial deterioration, in addition to outright failure. Broadening the outcome definition in this way raises event frequency to a level supporting estimation while retaining information about the rare outcome of ultimate concern.

This scoping is a substantive design decision rather than a hedge. Relative ordering requires monotonicity in exposure rather than calibration, a materially weaker condition, and it is sufficient for the uses to which the estimates are put: prioritizing assessment effort, triggering escalation at defined thresholds, and identifying concentration. Claiming calibrated absolute probabilities on the available outcome base would not be defensible, and a framework claiming them would fail at implementation.

Estimation approach: Structured expert elicitation supplies prior distributions where direct observation is insufficient, following methods developed for precisely these conditions (O'Hagan *et al.*, 2006), with the elicitation protocol documented so that the resulting priors are auditable rather than tacit. Observed adverse events update these priors as they accumulate. Hierarchical pooling across vendors sharing sector, size, service type, or jurisdiction extends the effective evidence base available to any individual vendor, so that a vendor with little individual history inherits information from comparable vendors while retaining the capacity to diverge as its own evidence accumulates.

The interpretive position follows Aven (2016): probabilities produced under these conditions express knowledge-conditional judgment rather than observed frequency, and the framework requires this to be stated in reporting rather than obscured by the appearance of numerical precision. Kaplan and Garrick's (1981) triplet formulation supplies the reporting structure, with scenario, likelihood, and consequence stated separately rather than collapsed into a single index.

AI role: Two functions. Signal-to-estimate mapping identifies how combinations of signals relate to adverse outcomes, including interactions that additive scoring cannot represent: simultaneous deterioration across financial and operational categories may carry different implications than the sum of the individual category movements. Anomaly detection identifies deviation from a vendor's own established baseline and from its peer group, surfacing change that absolute-threshold approaches miss.

Governance controls: The interagency model risk guidance (Federal Reserve and OCC, 2011) supplies the standard: models used in consequential decisions require validation against outcomes, documentation sufficient for independent review, and challenge by a function independent of the one operating them. The framework adopts these requirements. Validation compares estimates against realized adverse events on a defined schedule. Weighting and pooling parameters are documented and disclosed. Recalibration occurs on schedule and on trigger where validation reveals degradation. Challenge authority is vested outside the operating function, without which validation becomes self-assessment. Where model opacity would impede this challenge, Rudin's (2019) argument favours interpretable specifications in high-stakes settings, and the framework treats interpretability as a design constraint on Layer 2 rather than as a desirable property.

4.3. Layer 3: Governance architecture

Layer 3 addresses DR5 and instantiates the three conditions of governance anchoring defined in Section 2.3. The layer distinguishes this framework from treatments of risk quantification as a purely analytical problem, and it is the component that must answer the Power objection raised in Section 2.3.

Advance accountability assignment: An accountability matrix specifies, for each vendor criticality tier and exposure threshold, which role reviews the estimate, which approves the resulting classification, and which authorizes response

action. The matrix is fixed before estimates are produced. Assignment after the fact permits the accountability to be allocated to whoever is willing to accept the implication, which is the mechanism through which inconvenient findings are absorbed.

Advance threshold specification: Escalation triggers are defined as conditions on the estimates: absolute threshold exceedance, rate of change beyond a specified parameter, uncertainty interval width exceeding a specified bound, or concentration indicators above a defined level. Specification in advance constrains the discretion through which a deteriorating signal is characterized as not yet warranting attention. The uncertainty-width trigger is the least conventional of these: an estimate whose uncertainty is widening carries information even when its central value has not moved, and prevailing practice has no mechanism for acting on this.

Independent assurance of the analytical layers: The layer subjects Layers 1, 2, and 4 to scheduled assurance. For Layer 1 this covers signal source integrity, provenance completeness, and the independent-source proportion. For Layer 2 it covers model validation status, bias across vendor characteristics including jurisdiction and size, drift against changing conditions, and interpretability sufficient to support challenge. For Layer 4 it covers the dependence structure on which aggregation rests: whether dependency mapping remains current, whether the assumptions substituting for unobserved sub-tier relationships remain defensible, and whether cross-entity calibration parameters continue to hold. Layer 4 warrants assurance attention disproportionate to its apparent simplicity, because its outputs are the most consequential in the framework and rest on the least observable inputs. An aggregate exposure figure carries authority that its underlying dependence assumptions may not support, and unexamined aggregation is a more dangerous failure than unexamined estimation. The methodological basis for these activities is the internal algorithmic auditing framework of Raji *et al.* (2020), which specifies audit activities across an AI system's lifecycle, with documentation following model reporting standards (Mitchell *et al.*, 2019). Assurance findings return to Layers 1, 2, and 4 as the paths shown in Figure 2.

External assurance of Layer 3: The governance layer cannot assure itself without reproducing the circularity the framework exists to avoid. Assurance of Layer 3 must originate outside the framework, whether through internal audit, a board-level risk committee, or supervisory examination. This is not a gap in the design but a consequence of its own rule that challenge authority sits independent of the function being challenged, applied consistently to the governance layer itself. The layer's outputs are specified to support such external review: the accountability matrix, threshold definitions, override records, and assurance findings constitute the evidence an external reviewer requires to determine whether the governance apparatus operates substantively or ceremonially.

Override protocol: Human decision-makers may override model-derived classifications. The protocol requires recorded rationale, and overrides are monitored for pattern. Systematic override in one direction indicates either model inadequacy or governance failure, and distinguishing between them is an assurance function. Without pattern monitoring, override becomes an unexamined channel through which the quantification is neutralized while appearing to operate.

These four mechanisms constitute the framework's answer to the ceremonial-elaboration objection. Each creates a specific path by which the system's own performance becomes visible and contestable. Whether they suffice is assessed in Section 6.1 rather than assumed here.

4.4. Layer 4: Aggregation and escalation

Layer 4 addresses DR4, consolidating entity-level estimates into portfolio and group views that reveal exposure invisible at the individual vendor level. The layer is the framework's response to the multi-entity condition of Section 2.4.

Dependence representation: Aggregation that assumes independence between vendors would defeat the layer's purpose, since correlated failure is the exposure of interest. The framework specifies dependence-aware aggregation, and the specification is necessary because the choice materially affects the result. The dependence sources represented are shared sub-tier providers, identified through dependency mapping; common exposure factors including jurisdiction, sector, and infrastructure dependency; and contagion paths where one vendor's failure directly impairs another. Aggregation proceeds through scenario-based simulation over the dependence structure rather than through summation of marginal estimates. Where dependence structure is estimated rather than observed, the estimation approach and its assumptions are documented as an output of the layer, since an aggregate exposure figure is uninterpretable without them. The literature on dependence modelling in risk aggregation establishes both the methods and their failure modes, particularly the inadequacy of linear correlation for representing dependence in the tails where aggregation matters most (Embrechts *et al.*, 2002).

Figure 3 illustrates why the assumption is consequential rather than technical.

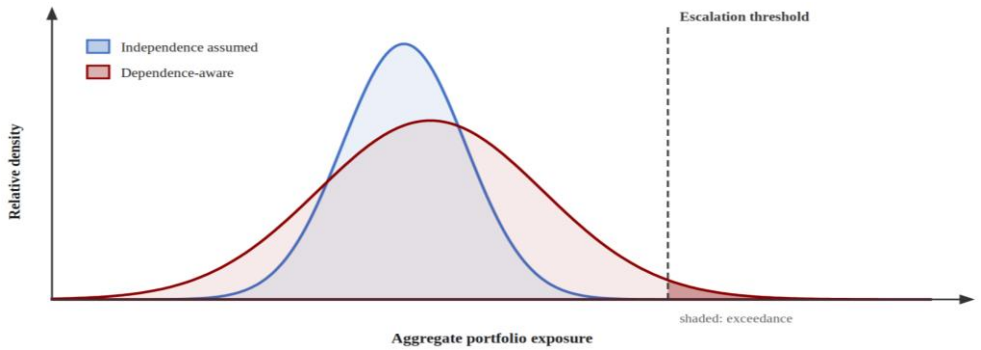

Fig 3; Schematic illustration of the effect of the dependence assumption on aggregate portfolio exposure. The distributions are illustrative rather than empirical, representing the property established in Embrechts *et al.* (2002)

The two aggregations diverge least at the centre of the distribution and most in the tail. This matters because escalation thresholds are set in the tail: an organization sets thresholds to detect the exposure levels it cannot absorb, not the levels it expects. Aggregation under an independence assumption therefore understates exceedance probability precisely in the region the threshold exists to monitor, and it does so while producing a central estimate that appears reasonable. The failure is not that the independent aggregation is obviously wrong but that it is plausibly wrong in the one region where being wrong is consequential.

Cross-entity consolidation: Entities applying different instruments and calibrations produce estimates that are not directly comparable. The layer specifies a calibration protocol converting entity-level estimates to a common scale before consolidation, with the conversion documented and its residual error reported. Consolidation without calibration produces a group figure that appears authoritative while aggregating incommensurable inputs, which is the failure mode of present practice rather than a solution to it.

Concentration detection: The layer computes concentration within an entity, across entities sharing a vendor, and at sub-tier level where multiple vendors depend on common providers. Sub-tier concentration is the exposure that individual assessment structurally cannot surface, and Section 5 demonstrates it in both cases examined.

Escalation: Group-level thresholds trigger reporting to group risk committees and, where supervisory obligations attach at group level, supply the consolidated evidence base those obligations require.

5. Demonstration

Demonstration applies the framework to two documented cases selected for contrasting failure modes: progressive financial and operational deterioration in one, and cybersecurity compromise in the other. The pairing exercises a wider portion of the framework than a single case would, and the contrast tests whether the design generalizes beyond one failure type.

Both cases rest on independent investigative records. Claims about each are sourced to those records, and where the record does not support a claim the framework's response is stated as a design property rather than as a counterfactual assertion. The limits of retrospective demonstration are addressed in Section 5.3.

5.1. Case 1: Carillion

Carillion plc, a UK construction and facilities management group employing approximately 43,000 people and holding extensive public sector contracts, entered compulsory liquidation on 15 January 2018. The collapse was investigated by the National Audit Office (2018) and jointly by two House of Commons select committees (House of Commons, 2018), producing a detailed public record of the preceding period.

The case is instructive because deterioration was progressive and substantially visible in public information. The investigative record documents that the company issued a profit warning in July 2017 disclosing contract provisions of approximately £845 million, followed by further warnings in September and November 2017; that the group carried substantial and growing debt alongside a significant pension deficit; and that the company's accounting treatment of contract revenue and its use of supply chain finance arrangements attracted external scrutiny before the collapse. The record further documents that Carillion had been among the most heavily shorted stocks on the London market for an extended period preceding failure, indicating that external market participants had formed adverse views on information publicly available. Notwithstanding the July 2017 warning, government bodies awarded Carillion further contracts in the subsequent period, a sequence the National Audit Office examined directly.

Framework response: Layer 1 ingests the categories of evidence the investigative record shows were available: filed accounts and interim statements, market indicators including short interest, credit assessments, and public reporting. Under prevailing periodic assessment, a vendor reviewed annually would carry a rating set before the July 2017 warning through the subsequent deterioration. Continuous ingestion registers each disclosure as it occurs. This is a property of the layer's design rather than a counterfactual claim about detection, and it addresses DR1 and DR2 directly.

Layer 2 registers progressive deterioration as movement in the exposure estimate and, equally importantly, as widening of the uncertainty interval. The composite adverse-event construct is material here: contract provisions, margin compression, and payment-term extension to subcontractors are frequent, observable events carrying information about the rare terminal outcome, and an estimation target restricted to failure alone would have had nothing to work with until failure occurred.

Layer 3 is where the case bears most directly on the framework's central argument. The investigative record documents that adverse information was available to counterparties and that contracting nonetheless continued. This is not a detection failure but an escalation failure, precisely the pattern Power (1997, 2007) describes: information existed within the system without acquiring decision force. The layer's response is the advance specification of thresholds and accountability, so that a rate-of-change trigger and a widening uncertainty interval produce a defined escalation to an identified role rather than an unallocated signal. Whether such specification would have altered the outcome cannot be established retrospectively. What the case establishes is that the failure mode the layer addresses is real and consequential, which is the relevant demonstration point.

Layer 4 addresses the exposure that entity-level assessment could not surface. Carillion held contracts across multiple government departments and public bodies, each contracting and assessing separately. No single assessing body held the consolidated view of aggregate public sector exposure to a single counterparty, and the sub-tier exposure through Carillion's subcontractor base extended further still. Cross-entity consolidation and shared-vendor concentration detection address this directly, and the case demonstrates DR4's practical significance more clearly than any single-entity example could.

5.2. Case 2: SolarWinds

In December 2020, the US Cybersecurity and Infrastructure Security Agency issued an alert documenting that malicious actors had compromised software updates distributed through the SolarWinds Orion platform, affecting a substantial population of government and private sector customers (CISA, 2020).

The case contrasts with Carillion in every relevant dimension.

Deterioration was not progressive but discrete. The compromise was not visible in financial or operational disclosure. The threat actor was sophisticated and, per the investigative record, state-associated.

Framework response, and its limits: The framework would not have detected this compromise, and claiming otherwise would misrepresent both the case and the artifact. No vendor risk assessment instrument operating on disclosure, financial, and market evidence detects a covert supply chain implant. Stating this is necessary because a demonstration that claims universal efficacy demonstrates nothing.

What the case does exercise is Layer 4. The compromise propagated through a vendor occupying a privileged position across a large population of dependent organizations, each of which assessed SolarWinds independently. The exposure was not the individual vendor relationship but the concentration: a single provider with privileged access across a dependent network. This is precisely the sub-tier and shared-dependency concentration that Layer 4 computes and that individual assessment structurally cannot surface. The case therefore demonstrates DR4 from a different direction than Carillion, showing that concentration exposure arises from privileged access position independently of any deterioration in the vendor's observable condition.

The case also bears on Layer 3's assurance function in a secondary sense. An organization's exposure to a privileged-access vendor is a governance parameter independent of that vendor's assessed condition, and the framework's threshold specification permits escalation on position rather than on estimate movement alone.

5.3. Demonstration summary and limits

Table 5 states the two cases against the framework layers.

Table 5: Demonstration summary.

Layer	Carillion (progressive financial and operational deterioration)	SolarWinds (covert cybersecurity compromise)	Requirement exercised
Layer 1	Continuous ingestion registers profit warnings, filed accounts, market indicators, and credit assessment as they appear, rather than at an annual review point	Limited relevance; the compromise was not visible in disclosure or market evidence	DR1, DR2
Layer 2	Composite adverse-event construct captures contract provisions, margin compression, and payment-term extension, which precede and inform the rare terminal outcome; uncertainty interval widens as evidence becomes conflicting	Not exercised; no signal available to the estimation layer	DR3, DR6
Layer 3	Addresses the escalation failure documented in the investigative record: advance threshold and accountability specification convert available information into a defined organizational response	Threshold specification on privileged-access position rather than on estimate movement	DR5
Layer 4	Cross-entity consolidation reveals aggregate public sector exposure to a single counterparty that no individual contracting body held	Shared-dependency concentration across a population of organizations each assessing the vendor independently	DR4

Retrospective application carries hindsight bias structurally. The analysis identifies signals whose significance is known because the outcome is known, and it cannot establish that the framework would have distinguished them prospectively from the volume of signals a portfolio generates. This limitation is inherent to retrospective case demonstration and is not remedied by additional cases.

The demonstration establishes that the failure mechanisms the framework addresses are real and documented, not that the framework would have prevented the outcomes. Carillion demonstrates an escalation failure of the kind Layer 3 targets. SolarWinds demonstrates a concentration exposure of the kind Layer 4 targets. Neither establishes efficacy.

Two cases, however contrasting, do not span the third-party risk domain. Regulatory non-compliance, geopolitical disruption, and gradual service degradation are unexercised.

6. Discussion

6.1. Evaluation against external referents

Table 6 reports the evaluation designed in Section 3.5, applying external referents to each design requirement.

Table 6: Design requirement evaluation against external referents

Requirement	Referents applied	Assessment	Basis and residual conditions
DR1: Continuity	Supervisory expectation (OCC 2013-29 ongoing monitoring; NIST SP 800-161); documented failure evidence (Carillion)	Adequately addressed	Layer 1 specifies continuous ingestion satisfying ongoing monitoring expectations, and the Carillion record confirms that the evidence such ingestion would capture existed publicly. Residual condition: continuous external evidence is unevenly available across vendor types, being substantially richer for listed than for private entities, and the framework does not resolve this asymmetry
DR2: Independence	Supervisory expectation (OCC 2013-29 risk-based due diligence); methodological standard (textual analysis literature)	Partially addressed	External evidence reduces dependence on vendor self-report, and the extraction methodology is established (Loughran and McDonald, 2011, 2016). It does not eliminate asymmetry: control environment, internal incident history, and sub-tier arrangements remain observable only through vendor cooperation. The framework improves the balance of evidence without closing the gap, and overstating this would misrepresent it
DR3: Quantification	Methodological standard (Cox, 2008; Kaplan and Garrick, 1981); supervisory expectation (ISO	Adequately addressed	Layer 2 produces continuous estimates with explicit uncertainty, satisfying the arithmetic combination requirement that Cox (2008) shows ordinal categories cannot meet, and reporting in the triplet structure of

	31000:2018)		Kaplan and Garrick (1981). Assessment is conditional on the DR6 scoping: the requirement is met for relative ordering, not for absolute calibration
DR4: Portfolio and group awareness	Documented failure evidence (both cases); methodological standard (dependence modelling literature)	Conditionally addressed	Layer 4 specifies dependence-aware aggregation and both cases confirm the exposure is real and consequential. Satisfaction depends on sub-tier dependency data that organizations frequently do not hold and that the framework does not supply a means of obtaining. The layer specifies what must be known without establishing how to know it, and this is the framework's most significant unresolved condition
DR5: Governance anchoring	Supervisory expectation (Federal Reserve and OCC, 2011 model risk guidance); methodological standard (Raji *et al*., 2020; Mitchell *et al*., 2019)	Adequately addressed	Layer 3 satisfies the three defined conditions and maps to the model risk guidance requirements for validation, documentation, and independent challenge, with assurance activities following established algorithmic audit methodology and extending to the dependence assumptions in Layer 4. The requirement that the governance layer itself be assured externally applies the independence rule consistently. Residual condition: the mechanisms are structural, and structures can be operated ceremonially. The Power objection is answered at the level of design, not of practice, and only naturalistic evaluation can establish whether the answer holds
DR6: Data-condition realism	Methodological standard (O'Hagan *et al*., 2006; Aven, 2016)	Adequately addressed	Layer 2 scopes its claim to relative ordering, broadens the outcome construct to raise event frequency, and specifies elicitation and hierarchical pooling as established approaches to sparse-outcome estimation. The requirement is met by constraining the claim rather than by overcoming the constraint, which is the appropriate response to a binding data condition

Of the six requirements, four are adequately addressed, one partially, and one conditionally, and the differentiation is substantive rather than presentational. DR2 is partial because information asymmetry is reduced rather than resolved, and no arrangement of external evidence sources fully substitutes for access the vendor controls. DR4 is conditional because the layer's operation depends on sub-tier dependency data whose acquisition lies outside the framework, which is a genuine gap rather than a matter of implementation detail.

The residual condition on DR5 warrants emphasis because it concerns the study's central claim. The framework's answer to Power's ceremonial-elaboration critique operates at the level of design: it specifies feedback paths, override visibility, and independent challenge that ceremonial systems lack. Design specification does not guarantee practice. An organization can implement every mechanism and operate all of them as ritual, and nothing in an analytically evaluated artifact can exclude this. The claim defensible on present evidence is that the framework makes ceremonial operation more difficult to sustain undetected, not that it prevents it.

6.2. Theoretical contribution

Following Gregor and Hevner (2013), the contribution is positioned as exaptation: extending known solutions to a new problem domain where the extension requires non-trivial modification. Quantification approaches developed in credit risk are extended to third-party risk, and two modifications distinguish the extension from mere transfer.

The quantification claim itself is modified first. Credit risk quantification operates on dense outcome data supporting calibration. Third-party risk does not, and the framework accordingly scopes its claim to relative ordering with explicit uncertainty, broadens the outcome construct to raise event frequency, and combines elicited priors with hierarchical pooling. This is not a weaker version of the credit risk approach but a differently structured one, adapted to a binding constraint the source domain does not face.

Governance is modified in turn. Credit risk models operate within mature model risk management regimes developed over decades. No comparable regime exists for third-party risk quantification, and the framework specifies one by drawing on model risk supervisory expectations (Federal Reserve and OCC, 2011) and algorithmic accountability methodology (Raji *et al*., 2020; Mitchell *et al*., 2019), adapting both to a domain where the quantified object is an external organization rather than an internal exposure.

The contributions reach three literatures. To supply chain risk management, the framework operationalizes the network perspective that scholarship has advocated for two decades (Christopher and Peck, 2004; Craighead *et al*., 2007) [6] but that assessment practice has not adopted, specifying in Layer 4 the mechanism through which dependence structure enters organizational risk processes. To enterprise risk governance, it addresses the gap that empirical ERM research has identified between framework adoption and outcome improvement (Bromiley *et al*., 2015; Arena *et al*., 2010) by specifying governance controls as constitutive components of the analytical apparatus rather than as arrangements surrounding it. To scholarship on AI in risk assessment, it supplies a domain-specific governance specification for a field whose accountability work has developed largely at the level of general principle (Floridi *et al*., 2018; Jobin *et al*., 2019).

These contributions are conceptual. The framework articulates a design logic that, on implementation, could address the identified limitations. Whether it does, and under what conditions, are questions the study raises rather than answers.

6.3. The multi-entity contribution

Multi-entity conditions are the dimension least addressed in the literature reviewed, and the framework's treatment of them stands apart accordingly.

Existing treatments of third-party risk assume a single assessing organization facing a vendor population. The multi-entity case differs structurally rather than in degree: risk appetite set at group level and applied at entity level, vendor relationships shared without defined ownership, concentration accumulating at a level no entity assesses, and supervisory obligations attaching where the evidence does

not sit. Layer 4's calibration protocol and cross-entity consolidation address the first and fourth. Its shared-vendor concentration detection addresses the third. The second, ownership of shared relationships, is an organizational design question the framework surfaces through the accountability matrix in Layer 3 without resolving, since it depends on structures outside the artifact's scope.

The Carillion case demonstrates the practical weight of this dimension. Aggregate public sector exposure to a single counterparty was distributed across contracting bodies none of which held the consolidated view, and the resulting exposure was visible only after realization.

6.4. Implementation Conditions and Adoption

Design science artifacts are assessed on utility, which requires attention to the conditions under which implementation is worthwhile. Continuous multi-source ingestion, probabilistic estimation with documented elicitation, independent model challenge, and dependence-aware aggregation carry substantial cost in licensing, integration, analytical capacity, and governance attention. For many organizations and most vendors, that cost exceeds the value of improved assessment.

Portfolio criticality distribution bears on that calculation more than portfolio size does: an organization with two thousand vendors of which thirty are critical should apply the framework to the thirty, and the framework's layered structure supports this tiering, since Layers 1 and 2 can operate at reduced intensity for lower tiers while Layer 4's concentration analysis requires portfolio-wide coverage at lower resolution. Entity structure determines the value of Layer 4, which contributes little in single-entity organizations and most in groups with shared vendor populations across entities. Regulatory exposure affects the calculus, since organizations subject to consolidated supervisory reporting obligations derive value from cross-entity consolidation that unregulated organizations do not.

For organizations considering AI-enabled assessment capability, governance design should proceed alongside analytical capability rather than after it: the accountability matrix and threshold specification of Layer 3 are cheap to design before implementation and expensive to retrofit, and an analytical capability deployed without them produces outputs with no defined pathway to action. And the assurance function requires capability that risk and audit functions may not hold, since reviewing a quantification model for bias, drift, and interpretability differs from reviewing a control environment. Organizations should treat this as a capability requirement rather than an incremental task for existing assurance staff.

7. Conclusion

The study addressed how AI-enabled capabilities can be integrated into governance-anchored frameworks to transform third-party risk assessment from ordinal classification to probabilistic exposure quantification in multi-entity supply chain environments. Through design science research, it developed a four-layer framework and evaluated it against external referents comprising supervisory expectations, documented failure evidence, and established methodological standards.

What distinguishes the framework is that governance controls are specified as constitutive components of the analytical apparatus rather than as arrangements added subsequently, so that its answer to the objection that risk apparatus proliferates ceremonially rests on feedback paths, override visibility, and independent challenge built into the design. Its quantification claim is likewise scoped to what sparse outcome data supports, producing relative exposure ordering with explicit uncertainty rather than calibrated absolute probabilities, which is the modification that makes the exaptation from credit risk defensible rather than aspirational.

Against the six design requirements, four emerged as adequately addressed, one partially, and one conditionally, and those falling short do so for identified reasons: information asymmetry is reduced rather than resolved, and dependence-aware aggregation requires sub-tier data the framework does not supply a means of obtaining.

8. Limitations

Empirical Validation is Absent: The framework has been evaluated analytically in an artificial setting, following the formative evaluation positioning of Venable *et al*. (2016). Demonstration on two documented cases establishes that the failure mechanisms addressed are real, not that the framework would have altered the outcomes. The relationship between adoption and assessment quality remains a design proposition.

The Knowledge Base Was Assembled Purposively: Sources were selected for their function in the design rather than through systematic search, as Section 3.3 states. The study therefore cannot claim that contrary evidence was sought and not found, and its assessment of what the literature addresses characterizes the work engaged rather than establishing an absence. This is a recognized property of design studies, whose warrant rests on design logic and external-referent performance rather than on literature completeness (Hevner, 2007), but it bounds the positioning claim in Section 2.6 and readers should weigh that claim accordingly. A systematic review of the intersection would be a distinct contribution and is identified below as a research direction.

The source boundary closes at the end of 2021: This is a scope condition adopted deliberately, and its consequences are material. The technical capability landscape for textual analysis changed substantially after the boundary, and the Layer 1 specification is correspondingly conservative relative to capabilities that subsequently became available. The supervisory landscape for third-party risk also developed, and the regulatory referents used in evaluation should be read as expectations current to the covered period rather than as a complete statement of present requirements. Compounding the preceding limitation, the positioning assessment in Section 2.6 cannot speak to subsequent scholarship at all. Re-examining both the positioning claim and the framework's technical assumptions against post-boundary developments is the first item of the research agenda below.

The Study Was Conducted By A Single Researcher: Requirements derivation, framework development, and evaluation were performed without inter-researcher validation, independent coding, or structured peer debriefing. The traceability chain running from Table 1 through Tables 3 and 6 permits others to assess each step independently, which partially offsets the constraint, but it does not substitute for triangulation during the design process. Single-researcher design carries particular risk in design science, where the same party defines the problem, builds the artifact, and evaluates it; the use of external evaluation referents in Section

3.5 was adopted specifically to mitigate this, and readers should assess whether the mitigation is sufficient.

Secondary Sources Only: No practitioner input informed the framework. Its practical feasibility, organizational acceptability, and implementation requirements have not been assessed by those who would operate it. The adoption conditions in Section 6.4 are reasoned rather than observed.

Two Cases, Retrospectively Applied: The demonstration exercises two failure modes and carries structural hindsight bias, as Section 5.3 states. Regulatory non-compliance, geopolitical disruption, and gradual service degradation remain unexercised.

Functional Rather Than Technical Specification: AI capabilities are described at the level of what they change about the assessment process. This keeps the framework independent of specific implementations but leaves substantial technical design between specification and operation.

Jurisdictional Concentration: The supervisory instruments underpinning the framework are predominantly US and international-standard sources. Layer 3 would require adaptation for jurisdictions with different supervisory expectations or different legal constraints on automated decision-making.

Future research

A systematic review of the intersection of AI capability and third-party risk governance, extending beyond the source boundary observed here, is the immediate priority among the directions these limitations open: it would establish on evidence what the present study can only assess within the literature it engages, and it conditions the value of the others. Naturalistic evaluation through pilot implementation would move the requirement assessments in Table 6 from analytical to empirical grounding and would test whether the Layer 3 mechanisms resist ceremonial operation in practice, which analytical evaluation cannot establish. Practitioner-engaged research would assess feasibility and surface implementation barriers that secondary sources cannot. And the sub-tier dependency data problem identified as the residual condition on DR4 is a substantive research question in its own right, since dependence-aware aggregation across any framework requires dependency data the field does not currently know how to obtain at scale.

References

1. Arena M, Arnaboldi M, Azzone G. The organizational dynamics of enterprise risk management. Accounting, Organizations and Society. 2010;35(7):659-675.
2. Aven T. Risk assessment and risk management: review of recent advances on their foundation. European Journal of Operational Research. 2016;253(1):1-13.
3. Bode C, Wagner SM. Structural drivers of upstream supply chain complexity and the frequency of supply chain disruptions. Journal of Operations Management. 2015;36:215-228.
4. Bromiley P, McShane M, Nair A, Rustambekov E. Enterprise risk management: review, critique, and research directions. Long Range Planning. 2015;48(4):265-276.
5. Chopra S, Sodhi MS. Managing risk to avoid supply-chain breakdown. MIT Sloan Management Review. 2004;46(1):53-61.
6. Christopher M, Peck H. Building the resilient supply chain. International Journal of Logistics Management. 2004;15(2):1-14.
7. Cybersecurity and Infrastructure Security Agency. Alert AA20-352A: Advanced Persistent Threat Compromise of Government Agencies, Critical Infrastructure, and Private Sector Organizations. Washington, DC: Cybersecurity and Infrastructure Security Agency; 2020.
8. Committee of Sponsoring Organizations of the Treadway Commission. Enterprise Risk Management—Integrating with Strategy and Performance. Durham, NC: Committee of Sponsoring Organizations of the Treadway Commission; 2017.
9. Cox LA Jr. What's wrong with risk matrices? Risk Analysis. 2008;28(2):497-512.
10. Craighead CW, Blackhurst J, Rungtusanatham MJ, Handfield RB. The severity of supply chain disruptions: design characteristics and mitigation capabilities. Decision Sciences. 2007;38(1):131-156.
11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019; 2019; Minneapolis, MN. p. 4171-4186.
12. Eisenhardt KM. Agency theory: an assessment and review. Academy of Management Review. 1989;14(1):57-74.
13. Embrechts P, McNeil A, Straumann D. Correlation and dependence in risk management: properties and pitfalls. In: Dempster MAH, editor. Risk Management: Value at Risk and Beyond. Cambridge: Cambridge University Press; 2002. p. 176-223.
14. Fan Y, Stevenson M. A review of supply chain risk management: definition, theory, and research agenda. International Journal of Physical Distribution and Logistics Management. 2018;48(3):205-230.
15. Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency. Supervisory Guidance on Model Risk Management: SR Letter 11-7 / OCC Bulletin 2011-12. Washington, DC: Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency; 2011.
16. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, *et al*. AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines. 2018;28(4):689-707.
17. Gregor S, Hevner AR. Positioning and presenting design science research for maximum impact. MIS Quarterly. 2013;37(2):337-355.
18. Gregor S, Jones D. The anatomy of a design theory. Journal of the Association for Information Systems. 2007;8(5):312-335.
19. Hevner AR. A three cycle view of design science research. Scandinavian Journal of Information Systems. 2007;19(2):87-92.
20. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Quarterly. 2004;28(1):75-105.
21. Ho W, Zheng T, Yildiz H, Talluri S. Supply chain risk management: a literature review. International Journal of Production Research. 2015;53(16):5031-5069.
22. House of Commons. Carillion: Second Joint Report from the Business, Energy and Industrial Strategy and Work and Pensions Committees of Session 2017-19. HC 769. London: House of Commons; 2018.

23. International Organization for Standardization. Risk Management—Guidelines. ISO 31000. Geneva: International Organization for Standardization; 2018.
24. Ivanov D. Predicting the impacts of epidemic outbreaks on global supply chains: a simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. Transportation Research Part E: Logistics and Transportation Review. 2020;136:101922.
25. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nature Machine Intelligence. 2019;1(9):389-399.
26. Kaplan RS, Mikes A. Managing risks: a new framework. Harvard Business Review. 2012;90(6):48-60.
27. Kaplan S, Garrick BJ. On the quantitative definition of risk. Risk Analysis. 1981;1(1):11-27.
28. Khandani AE, Kim AJ, Lo AW. Consumer credit-risk models via machine-learning algorithms. Journal of Banking and Finance. 2010;34(11):2767-2787.
29. Leo M, Sharma S, Maddulety K. Machine learning in banking risk management: a literature review. Risks. 2019;7(1):29.
30. Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance. 2011;66(1):35-65.
31. Loughran T, McDonald B. Textual analysis in accounting and finance: a survey. Journal of Accounting Research. 2016;54(4):1187-1230.
32. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, *et al*. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency; 2019; Atlanta, GA. p. 220-229.
33. National Audit Office. Investigation into the Government's Handling of the Collapse of Carillion. HC 1002. London: National Audit Office; 2018.
34. Ngai EWT, Hu Y, Wong YH, Chen Y, Sun X. The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. Decision Support Systems. 2011;50(3):559-569.
35. National Institute of Standards and Technology. Supply Chain Risk Management Practices for Federal Information Systems and Organizations. SP 800-161. Gaithersburg, MD: National Institute of Standards and Technology; 2015.
36. Office of the Comptroller of the Currency. Third-Party Relationships: Risk Management Guidance. Bulletin 2013-29. Washington, DC: Office of the Comptroller of the Currency; 2013.
37. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, *et al*. Uncertain Judgements: Eliciting Experts' Probabilities. Chichester: Wiley; 2006.
38. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. Journal of Management Information Systems. 2007;24(3):45-77.
39. Power M. The Audit Society: Rituals of Verification. Oxford: Oxford University Press; 1997.
40. Power M. Organized Uncertainty: Designing a World of Risk Management. Oxford: Oxford University Press; 2007.
41. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, *et al*. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; 2020; Barcelona. p. 33-44.
42. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1:206-215.
43. Venable J, Pries-Heje J, Baskerville R. FEDS: a framework for evaluation in design science research. European Journal of Information Systems. 2016;25(1):77-89.
44. Zsidisin GA. A grounded definition of supply risk. Journal of Purchasing and Supply Management. 2003;9(5-6):217-224.