SUPPLY CHAIN RESEARCH · EDITORIAL FEATURE

The Supplier Risk Score Illusion

What Ratings and Audits Actually Measure, and What They Reliably Miss

Two raters assessing the same supplier reach different answers more often than they agree. Audited factories collapsed. Audited companies failed. The score is a triage device that the market reads as a prediction, and the gap between those two things is where the losses happen.

Independent, vendor-neutral research for supply chain and technology leaders

Supply Chain Research | supplychainresearch.com | 2026

Contents

What Ratings and Audits Actually Measure, and What They Reliably Miss

01 Executive summary

A modern supplier risk programme produces numbers. Each supplier receives a risk score, an environmental and social rating, an audit result, and a status in a monitoring platform, and those numbers flow into dashboards, onboarding decisions, category reviews, and board reporting. The apparatus is expensive, it is staffed by capable people, and it produces a comforting artefact: a portfolio in which every supplier has been assessed and most of them are green. The difficulty is that the numbers are being asked to do something they were not built to do. A score is a triage instrument, a device for ranking where to look first, assembled from what happened to be visible and documented at the time of assessment. It is routinely read as a prediction of whether a supplier will fail, and as a reason to stop looking.

The evidence for the gap between those two readings is substantial and comes from peer-reviewed work rather than from vendor marketing. Analysis of six major rating providers found that their assessments of the same companies correlate between roughly zero point three eight and zero point seven one, averaging about zero point six one, against roughly zero point nine or above for credit ratings from the major agencies. When the divergence is decomposed, the largest single contributor is measurement, meaning that raters looking at the same category on the same firm reach different conclusions about it, rather than that they value different things. Alongside this sit the cases: a factory audited under a social-compliance scheme months before it collapsed with more than eleven hundred deaths, a company that received nineteen consecutive unqualified audit opinions before entering liquidation with billions in liabilities, and a finance business with investment-grade backing whose collapse froze ten billion dollars of funds. We say honestly that no screening at all is worse than imperfect screening, and this article does not argue for abandoning supplier assessment. It argues for using it as triage, supplementing it with signals scores cannot capture, preferring assurance whose incentives are not compromised, extending assessment to the tier where failures actually originate, and keeping a record of what the model missed.

0.38 to 0.71
the range of pairwise correlation between six major rating providers assessing the same firms

56%
of rating divergence attributable to measurement rather than to differing scope or weights

19
consecutive unqualified audit opinions issued to a company that then collapsed with roughly GBP 7 billion of liabilities

The argument in brief

- **A score is triage, not prediction.** It ranks where to look and records what was documented on the day. Used as a verdict, it becomes a reason to stop looking at exactly the suppliers that most warrant attention.
- **Raters assessing the same firm disagree.** Pairwise correlations averaging around 0.61, against roughly 0.90 for credit ratings, mean the construct is far less settled than the single-number output implies.

- **The disagreement is mostly measurement.** The comfortable explanation, that raters simply weight different values, accounts for the smallest share of divergence. Most of it is raters measuring the same thing differently.
- **Assurance follows the money.** Research on supply chain monitoring finds third-party audits less effective than buyers' own, and points to who pays the auditor as the explanation.
- **Assessment concentrates where failures do not originate.** Programmes cover direct suppliers thoroughly and the sub-tiers barely at all, while a substantial share of disruption arrives from below tier one.

02 Triage device, read as a verdict

The central claim of this article is a claim about interpretation rather than about arithmetic, and it is worth stating precisely before the evidence is assembled, because the distinction it rests on is easy to nod at and hard to maintain in practice. Figure 1 sets it out.

Figure 1. The category error at the centre of supplier risk scoring. Scores are triage instruments that rank where to look and record what was visible. They are routinely used as predictions of failure and as reasons to stop looking.

Consider what a supplier risk score actually contains. It is assembled from inputs that exist in retrievable form: disclosed policies, held certifications, financial statements where available, public adverse-media signals, questionnaire responses, and audit results. Each of those is a record of something documented at a point in time. The score aggregates them according to a weighting the provider has chosen, and produces a figure that ranks this supplier against others assessed by the same method. That is a coherent and useful object. It permits a procurement function with four thousand suppliers and capacity to examine forty of them properly to decide which forty, which is a real problem that the score truly solves.

What the score cannot be is a statement about whether this supplier will fail, because the inputs do not contain the information that determines failure. A supplier fails because of financial stress that has not yet surfaced in published accounts, because of a concentration in its own supply base that nobody outside it can see, because of management decisions taken after the assessment, or because of practices that are deliberately concealed from assessors. None of that is in the retrievable record on the day the score is computed, and no weighting of the retrievable record produces it. The score summarises the visible; failure originates disproportionately in the invisible.

The category error occurs when the score is used to close inquiry rather than to direct it. A supplier that scores well is treated as assessed, moves into the approved population, and receives no further examination until its annual refresh. The organization has converted an absence of visible problems into a positive finding of soundness, which is a different proposition entirely, and it has done so at the exact moment when a supplier that is concealing something has most successfully achieved its objective. The best-concealed problems produce the cleanest scores, which means the score is weakest precisely where its failure is most costly.

This is not an argument that the scores are worthless, and readers who take it that way will misapply everything that follows. It is an argument about what class of instrument they are. A metal detector at an airport is a triage device: it directs attention, it produces false positives and false negatives, and nobody imagines that passing through it establishes a traveller's intentions. Supplier risk scores occupy

the same category and are frequently accorded a different status, appearing in board reporting as though the proportion of suppliers rated green were a measure of exposure rather than a measure of assessment coverage. The sections that follow examine how far apart those two things are.

There is an organizational dynamic that entrenches the category error and deserves naming, because it explains why the correct interpretation is so hard to sustain even where individuals understand it. A supplier risk function is accountable for coverage, which is measurable, and cannot be accountable for prediction, which is not. Its performance is therefore assessed on how many suppliers have been assessed, how current the assessments are, and how complete the files look, all of which are activity measures. Nobody in the reporting chain is rewarded for saying that a green rating on a critical sole-source supplier tells the organization considerably less than it appears to.

The result is a reporting convention in which assessment coverage is presented as though it were a risk position. A board paper stating that ninety-four percent of suppliers have been assessed and eighty-one percent are rated low risk conveys, to a reader who has not thought carefully about the instrument, that the supply base is in good order. It conveys, accurately, that the function has completed its process and that most suppliers had nothing visibly wrong on the day they were examined. Those are different statements, the second is much weaker, and the language of the reporting does not distinguish them.

03 Raters looking at the same firm disagree

The most direct evidence that these measures are less settled than their presentation suggests comes from comparing what different providers say about the same companies. Figure 2 presents the finding.

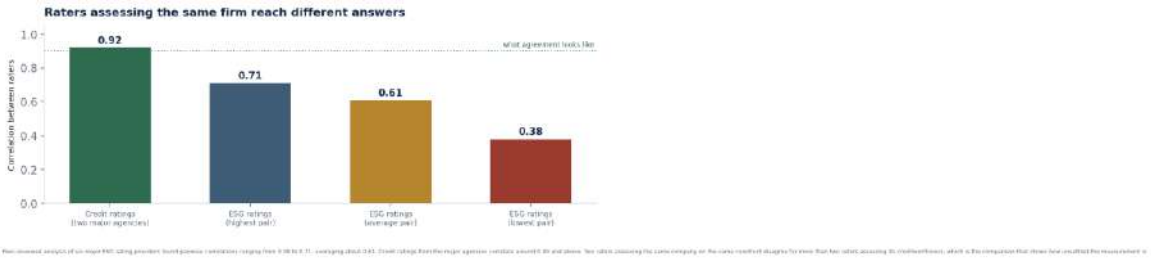

Figure 2. Peer-reviewed analysis of six major rating providers found pairwise correlations from 0.38 to 0.71, averaging about 0.61. Credit ratings from the major agencies correlate around 0.90 and above.

The study examined ratings from six prominent providers covering environmental, social, and governance performance, and computed how closely each pair of providers agreed in their assessments of the same firms. The correlations ranged from about zero point three eight at the lowest pair to about zero point seven one at the highest, with an average across pairs of roughly zero point six one. To interpret those figures, the natural comparison is credit ratings, where the major agencies assessing the same issuers correlate at roughly zero point nine or above. Credit rating is not a perfect science and the agencies do disagree, but they disagree within a much narrower band, and the gap between the two fields is the finding that matters.

The practical implication for a buyer is uncomfortable and immediate. If an organization screens its supply base using one provider's ratings and its competitor screens using another's, the two will identify substantially different sets of high-risk suppliers from the same population. A supplier excluded by one buyer as unacceptable may be approved by another as sound, and neither buyer is being careless. The variation is a property of the measurement rather than of the suppliers, which means that a decision to disqualify a supplier on the basis of a single provider's rating is a decision that would have gone differently had the organization subscribed elsewhere.

It also means that the confidence conveyed by the presentation of these ratings is not supported by their reproducibility. A rating expressed as a letter grade or a two-digit number carries an implicit claim to precision, and it is consumed that way: procurement functions set thresholds, exclude below a cut-off, and report the distribution to boards as though the categories were stable. A measure whose average inter-rater correlation is around zero point six does not support threshold decisions at that granularity, and using it that way produces the appearance of rigour without its substance. The appropriate use of such a measure is to rank and to flag, not to draw lines.

A further consequence of inter-rater divergence is worth drawing out because it affects suppliers as well as buyers. A supplier subject to assessment by several of its customers, each using a different provider, receives materially different assessments of itself and is asked to remediate different things by each. Since remediation is costly, and since the ratings do not agree on what is deficient, the supplier faces a choice between satisfying everyone, which is expensive and incoherent, and optimising for whichever

rating its largest customer uses. The predictable result is that suppliers manage to the metric rather than to the underlying condition, which is the standard response to any measurement regime with weak construct validity.

That response degrades the measure further over time. A rating that suppliers actively manage becomes progressively less informative about the thing it was meant to capture, because effort flows to the observable proxies rather than to the underlying practice, and the correlation between proxy and practice weakens. This is a familiar dynamic wherever a measure becomes a target, and supplier assessment is unusually exposed to it because the observables, policies and certifications and questionnaire responses, are cheap to improve relative to the practices they are meant to indicate.

04 Why they disagree, and what that implies

The reason the raters diverge matters more than the fact of divergence, because the two available explanations have sharply different implications. Figure 3 shows the decomposition.

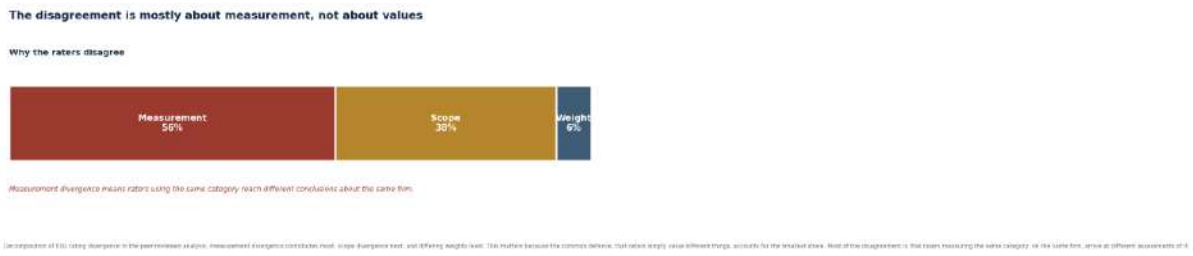

Figure 3. Decomposition of rating divergence: measurement contributes most, scope next, and differing weights least. The common defence, that raters simply value different things, accounts for the smallest share.

The comfortable explanation for divergence is that raters have different priorities, and that this is a feature rather than a defect. On this account, one provider weights carbon intensity heavily while another emphasises labour practices, so their rankings differ because they are answering different questions, and a buyer should simply choose the provider whose priorities match its own. If that were the whole story, the divergence would be unproblematic and the remedy would be to read the methodology and pick accordingly.

The decomposition in the study does not support that account. Divergence attributable to differing weights, the values explanation, contributes the smallest share of the total. The largest contributor is measurement divergence, meaning that raters assessing the same category on the same firm arrive at different assessments of that category, and the second is scope, meaning that they include different attributes in the assessment. Measurement divergence is a different kind of problem entirely, because it cannot be resolved by choosing a provider whose values you share: the providers are attempting to measure the same thing and getting different answers.

A concrete illustration makes the mechanism clear. Suppose two providers both assess labour practices. One uses disclosed policy documents and certification status; the other uses incident reports, litigation records, and workforce turnover. Both are measuring labour practices, both approaches are defensible, and they will produce different assessments of the same firm, because a company with excellent documented policies and poor actual practice scores well on the first and badly on the second. Neither provider is wrong. The construct they are both trying to measure is not directly observable, each has chosen a set of observable proxies, and the proxies do not agree.

The implication for a buyer is that these ratings should be treated as estimates with substantial uncertainty rather than as measurements, and that a supplier's rating should be understood as one estimate from one method rather than as its risk profile. Where a decision is consequential, the sensible response is to obtain more than one view and to examine the disagreement, because the cases where providers diverge sharply are informative in themselves: they indicate a supplier whose profile depends heavily on which observables are examined, which is a signal worth investigating rather than averaging away.

A practical test can settle for any organization whether this analysis applies to its own programme, and it takes an afternoon. Select a dozen suppliers spread across the risk distribution and obtain assessments of each from a second provider, then compare. Where the two agree, the assessment can be relied on to the modest extent this article allows. Where they diverge sharply, examine why, and the answer will indicate which observables each provider is weighting and what each is missing. The exercise costs little, produces immediate calibration of an instrument the organization is already using extensively, and is almost never performed.

05 Nineteen unqualified opinions

The abstract point about measurement uncertainty acquires force from the cases in which formal assurance signals remained favourable up to the point of collapse. The clearest documented example concerns a company that was, on every visible indicator, adequately assured. Figure 4 sets out the sequence.

Figure 4. A major construction and services group received nineteen consecutive unqualified audit opinions, announced a substantial contract write-down in July 2017, and entered compulsory liquidation in January 2018.

The company was one of the largest construction and support-services groups in its market and a strategic supplier to public bodies, delivering facilities management, infrastructure, and services under long-term contracts. Its accounts were audited by one of the largest firms in the profession, which issued unqualified opinions for nineteen consecutive years. In July 2017 the company announced a contract write-down of roughly eight hundred and forty-five million pounds. In January 2018 it entered compulsory liquidation with approximately seven billion pounds of liabilities against roughly twenty-nine million pounds of cash. The financial reporting regulator subsequently fined the audit firm, with the penalty reduced on settlement, and the national audit office estimated a substantial direct cost to taxpayers from the collapse.

It is important to be careful about what this case demonstrates, because the temptation is to draw too strong a conclusion. It does not establish that audits are worthless, that the auditors should have predicted the timing, or that any assessment methodology could have identified the failure in advance with confidence. Audit is designed to express an opinion on whether financial statements give a true and fair view, not to forecast solvency, and that distinction is real and frequently misunderstood by the people who consume audit opinions.

What it does demonstrate is that the formal assurance signals available to a buyer carried no warning. An organization performing supplier due diligence on this company at any point before mid-2017 would have found audited accounts with clean opinions, substantial public-sector contract wins, and a strategic supplier designation. Every retrievable indicator was favourable, and the score any reasonable methodology would have produced from those inputs would have been reassuring. The failure was not that the assessment was done badly; it was that the assessment could only see what was documented, and what was documented did not contain the problem.

The lesson generalises directly to supplier risk scoring, which draws on the same class of inputs and therefore inherits the same blind spot. A scoring model that ingests audited financial statements, certifications, and disclosures is ingesting the outputs of assurance processes with known and documented limitations, and it is aggregating them into a figure that presents with more confidence

than any of its inputs individually warrant. The aggregation does not reduce the uncertainty in the underlying signals; it conceals it behind a number.

06 What the scores said, and what happened

The audit case is one instance of a pattern that recurs across different assessment regimes and different kinds of failure. Figure 5 assembles the cases that this article relies on, with the caution that they are existence proofs rather than a sample.

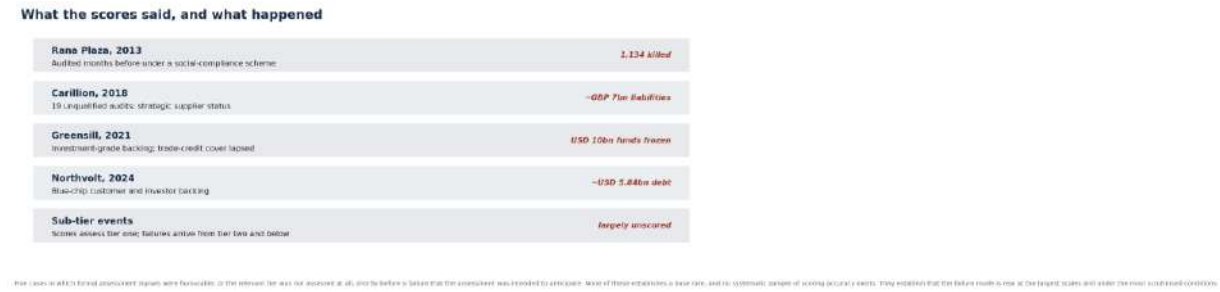

Figure 5. Five cases in which formal assessment signals were favourable, or the relevant tier was not assessed at all, shortly before a failure the assessment was intended to anticipate.

The most consequential is the 2013 collapse of a garment factory complex in Bangladesh in which one thousand one hundred and thirty-four people were killed. Documentation established afterward that one of the manufacturers operating in the building had been audited under a major social-compliance scheme in the months before the collapse. The audits were not structured to assess structural integrity of the building, which was the cause of the deaths, and this is a defence the audit regime has offered and which is technically accurate. It is also precisely the point: buyers relied on social-compliance audits as assurance about their supply chain, the audits assessed what they were scoped to assess, and the gap between what was assessed and what mattered was not visible to the buyers consuming the result.

The finance case is instructive for a different reason. A supply-chain finance business that had attracted investment from prominent backers and operated with substantial trade-credit insurance entered insolvency in March 2021 after that insurance cover lapsed, freezing approximately ten billion dollars in associated investment funds and precipitating distress at a large industrial group that depended on its financing. The chain of dependency ran from an insurer's decision, through a finance provider, to the working capital of manufacturing operations, and almost none of the parties exposed to that chain had it mapped. A supplier risk score assessing the manufacturing companies would have been assessing entities whose viability depended on a financing arrangement several steps removed from anything the score examined.

The battery manufacturer that entered bankruptcy in late 2024 with roughly five and eight tenths billion dollars of debt is the fourth case, and it is included because favourable assessment signals of a particular kind, blue-chip customer commitments and prominent investor backing, are frequently treated by buyers as a proxy for supplier soundness. They are a signal about what sophisticated parties believed, which is worth something and is not the same as a signal about whether the business will survive. The fifth entry is not a single case but a structural observation examined later: that assessment programmes concentrate on direct suppliers while a substantial share of disruption originates below them.

These cases establish that the failure mode is real at the largest scales and under the most scrutinised conditions, involving competent organizations with professional advisers and formal assurance in place. They do not establish a rate, and this article does not claim one, because no systematic sample of scoring accuracy exists. What they justify is a change in how the outputs are interpreted, which is the subject of the constructive sections.

A common and reasonable objection to this collection of cases is that each involves a different assessment regime, and that criticising financial audit, social-compliance audit, credit assessment, and investor diligence together conflates instruments designed for different purposes. The objection is correct as a matter of taxonomy and it does not weaken the argument, because the point being made is not about any single instrument. It is about how the outputs of all of them are consumed by buyers, who aggregate audited accounts, certifications, audit results, and investor signals into a general impression of supplier soundness that none of the individual instruments supports.

That aggregation is where the failure occurs. Each instrument, read within its own scope and with its limitations understood, is doing something defensible. A financial audit expresses an opinion on financial statements. A social-compliance audit assesses defined criteria at a site on a date. An investor's participation reflects that investor's judgment about return. A buyer combining these into a supplier risk rating has performed an aggregation nobody designed, validated, or takes responsibility for, and has produced from it a confidence that none of the inputs carries. The cases in this section are what that aggregation looks like when it fails.

07 Who pays the auditor

Audits sit underneath most supplier assessment regimes as the source of primary evidence, and there is a body of academic work examining whether they function as intended. One finding in particular runs against intuition and deserves attention from anyone designing an assessment programme. Figure 6 sets out the distinction.

Figure 6. Research on global supply chain monitoring has examined whether audits conducted by buyers' own staff differ from those conducted by independent third parties, and finds the third-party audits the less effective of the two.

The intuitive expectation is that independent third-party audits should be more rigorous than audits conducted by the buyer's own staff, because independence removes the auditor from the commercial relationship and should reduce the incentive to produce a favourable result. Academic work examining large samples of supply chain audits has found the opposite pattern, with third-party audits performing less well on measures of effectiveness than second-party audits conducted by the buying organization itself, and the proposed explanation concerns the structure of the commercial relationship rather than the competence or integrity of the auditors.

The mechanism is straightforward once stated. In many social-compliance regimes the audited supplier selects and pays the auditing firm, which means the auditor is competing for repeat engagement with the party it is assessing. An auditor that produces uncomfortable findings risks not being re-engaged, and an auditor that produces smooth results is easier to work with. No individual auditor need behave improperly for this to shape outcomes in aggregate, because the selection operates at the level of which firms get hired repeatedly. A buyer's own audit team faces no such pressure, because its employment does not depend on the audited party's satisfaction, and that difference in incentive appears to dominate the advantage that formal independence should confer.

Related work has examined what improves audit quality, and the practical levers identified include rotating auditors so that no individual or firm develops a settled relationship with a site, increasing the proportion of second-party audits, and using unannounced visits rather than scheduled ones. The last is the most direct: an announced audit assesses a site that has been prepared for assessment, which is a different thing from the site as it normally operates, and the difference is precisely the variable the audit is meant to capture. Buyers designing assessment programmes should treat announced third-party audits paid for by the audited party as the weakest available form of assurance, and should understand that a supplier file consisting entirely of such audits contains less information than its volume suggests.

The design of the assurance regime matters as much as the choice of auditor, and one variable dominates the others: whether the visit is announced. An announced audit assesses a site that has had notice to prepare, and preparation is rational and universal, covering documentation, housekeeping,

staffing on the day, and in some documented cases the temporary removal of conditions the audit would find. The audit then records the prepared state accurately, and the record enters the buyer's file as evidence about normal operation, which it is not. Unannounced visits are more expensive and less convenient, and they assess the thing the buyer actually wants to know about.

Related design choices compound in the same direction. Auditors who return to the same site repeatedly develop relationships that reduce the friction of the visit and the likelihood of adverse findings, which is why rotation is recommended. Audits scoped to a checklist find checklist items and miss anything outside the scope, which is how a structural hazard escapes a labour-practices audit. And audits whose findings carry no consequence produce compliance with the audit rather than with the standard. A buyer designing a regime should assume that every one of these variables will be optimised against by the audited party, not through bad faith but through the ordinary response of any organization to being measured, and should design accordingly.

08 What a score can see, and what it cannot

Underlying all of the preceding is a constraint that is structural rather than remediable: a scoring model can only use inputs that exist in retrievable form, and the variables that determine supplier failure are disproportionately not in that set. Figure 7 sets out the division.

Figure 7. Scores measure what is documented, not what happens. Policies, certifications, and audit results are retrievable; day-to-day practice, undisclosed subcontracting, and sub-tier financial stress generally are not.

On the visible side are the inputs every provider uses. Disclosed policy documents are retrievable and easy to score, and they appear in nearly every model. Certifications are binary, verifiable, and widely weighted. Audit results are point observations that enter the record as facts. Financial statements, where the supplier is required to file them, provide a lagged and audited view of the position at a past date. Public adverse-media signals capture what has already surfaced. Each of these is a legitimate input and each is a record of something that has been documented.

On the invisible side are the variables that actually drive outcomes. Actual practice between audits is not observed by anyone external, and the divergence between documented policy and daily practice is the specific thing social-compliance regimes exist to detect and demonstrably struggle to detect. Financial stress at a supplier's own suppliers is rarely assessed and frequently decisive. Undisclosed subcontracting defeats site-based methodology entirely, because the audited site is not where the work is done. And deliberate concealment is, by construction, invisible to announced assessment, which is why the best-concealed problems produce the cleanest files.

The consequence is that a supplier risk score is substantially a measure of disclosure quality and administrative maturity rather than of conduct or resilience. A supplier with a professional compliance function, well-drafted policies, current certifications, and a practised approach to audits will score well, and those attributes are correlated with being a well-run organization, which is why the scores have some predictive value. They are also exactly the attributes a supplier can develop without changing its actual practices, which is why the correlation is weaker than the presentation implies and why it breaks down at the suppliers where it matters most.

Recognising this changes what the buyer should do with the output rather than whether it should use it. The score identifies suppliers whose documentation is weak, which is a real signal, and it fails to identify suppliers whose documentation is strong and whose practice is not. The correct inference from a good score is that this supplier is administratively competent and nothing visible is wrong, which is plainly useful and considerably narrower than what the score is usually taken to mean. The correct response to

a good score at a supplier that matters is to obtain the kinds of evidence the score cannot contain, which is the subject of the protocol later in this article.

09 Assessing the tier that does not fail

A second structural limitation compounds the first, and it concerns where assessment effort is directed relative to where failures originate. Figure 8 shows the mismatch.

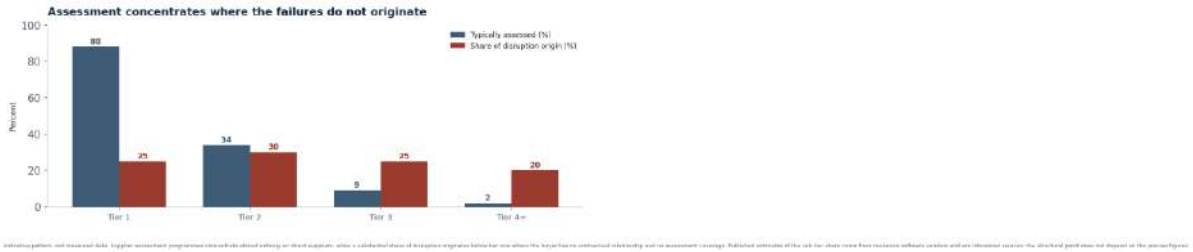

Figure 8. Indicative pattern, not measured data. Assessment programmes concentrate almost entirely on direct suppliers, while a substantial share of disruption originates below tier one where the buyer has no relationship and no coverage.

The reason for the concentration is not negligence but contract. A buyer has a commercial relationship with its direct suppliers and can require assessment, questionnaires, audits, and disclosure as a condition of doing business. It has no relationship with its suppliers' suppliers, no contractual right to demand anything of them, and frequently no knowledge of who they are, because that information is commercially sensitive and suppliers have reasons not to share it. Assessment therefore concentrates where the leverage exists, which is exactly where the buyer already has the most visibility and the least need.

The disruptions that have most damaged supply chains in recent years originated disproportionately below the first tier: single-source components several steps removed, specialised materials with concentrated production, contract manufacturers whose own capacity was constrained, and financial failures at parties the buyer had never heard of. Published estimates of the sub-tier share of disruption events come predominantly from vendors of supply chain resilience software, which is an interested source and is flagged as such, and the estimates vary by methodology. The structural point does not depend on the precise figure: a buyer that assesses tier one thoroughly and tier two barely has allocated its assessment effort in inverse proportion to where its ignorance lies.

Regulatory developments have begun to push in the other direction, requiring buyers in some jurisdictions to conduct due diligence extending beyond direct suppliers for particular risks, and import enforcement regimes have made sub-tier provenance a compliance question with direct commercial consequences. These developments create an obligation without creating the visibility, which is why the practical answer is usually not to attempt full multi-tier assessment, an exercise most organizations cannot resource, but to map dependency selectively: identify the components, materials, or capabilities on which the operation truly depends, trace those specific paths as deep as the dependency runs, and accept lower coverage elsewhere. Depth where it matters is achievable; uniform breadth is not, and a programme that promises uniform breadth is describing an aspiration rather than a practice.

Sub-tier assessment has a practical entry point that many organizations overlook, which is that the information frequently exists somewhere in the buyer's own records without having been assembled. Engineering specifications name approved component sources. Quality investigations trace defects to

their origin. Customs and origin documentation identifies where goods were produced. Logistics records show where shipments actually come from. None of this constitutes a supplier map, and together it contains a substantial part of one, held across functions that have no reason to combine it. Assembling what already exists is cheaper than any external mapping subscription and frequently reveals the concentrations that matter.

Where internal records are insufficient, the next step is to ask direct suppliers, which works better than buyers expect if the request is scoped narrowly. A general demand for full sub-tier disclosure meets resistance, because the information is commercially sensitive and the request appears open-ended. A specific question about who supplies a named critical component, framed as continuity planning rather than as audit, is answered far more often. The difference is that the narrow question is proportionate and the supplier can see why it is being asked, which is a general principle worth applying across supplier information requests.

10 The base rates nobody has measured

A reader persuaded by the argument so far will reasonably ask how often scores fail, and the answer is that nobody knows, which is itself a substantial finding about the state of this field.

What would be needed to answer the question is a study that took a large population of scored suppliers, followed them for a period, recorded which ones failed in ways the scoring was meant to anticipate, and computed the predictive performance of the scores against those outcomes. That study does not exist in the public literature so far as this research could establish. What exists instead is a large volume of material published by the firms that sell scoring, monitoring, and resilience products, describing the prevalence of disruption and the benefits of assessment, and that material is produced by parties whose commercial interest lies in establishing both that the problem is severe and that their instrument addresses it.

This is not an accusation of dishonesty and much of the vendor research is carefully constructed. It is an observation about what can be concluded from it. A figure describing the proportion of disruptions originating below tier one, published by a company selling multi-tier mapping software, may well be accurate and cannot be verified independently, because the underlying event population is the vendor's own customer base observed through the vendor's own instrument. The same applies to figures on how many suppliers experience distress annually, on the cost of disruption, and on the effectiveness of monitoring. Each is plausible, each is unverifiable, and each is repeated until its provenance disappears.

The absence of independent measurement has a practical consequence for buyers beyond epistemics. Without base rates, an organization cannot calibrate its own programme: it cannot say whether its supplier failure rate is high or low relative to peers, whether its assessment is catching more or fewer problems than a reasonable benchmark, or whether an increase in assessment spending would improve outcomes. It is operating a control system with no feedback about whether the control works. That is why the final section of the constructive material in this article recommends that organizations generate their own base rates from their own experience, which is the only measurement available to them and which almost none of them currently records.

There is a structural reason independent measurement has not appeared, beyond the absence of anyone with an incentive to fund it. The outcome variable is hard to define: supplier failure encompasses insolvency, quality collapse, delivery breakdown, compliance breach, and reputational event, which have different base rates and different relationships to any given assessment input. The population is hard to construct, because scoring coverage is not random and the suppliers that get assessed differ systematically from those that do not. And the data are held privately by parties with no obligation to share them. These are real obstacles rather than excuses, and they explain why the gap persists.

They also indicate why the organization's own record is the practical substitute. A single buyer can define failure in terms that matter to it, observe its own full supplier population rather than a selected sample, and hold outcome data that no external researcher can access. The resulting analysis is not generalisable and does not need to be, because the question being answered is about this organization's

instruments applied to this organization's supply base. What is unobtainable as public research is straightforward as private practice, which is an unusual and encouraging position to be in.

11 The fairness case: screening beats no screening

This article has argued hard against a set of practices that most large organizations follow, and the case in their favour is strong enough that a reader who concluded supplier assessment should be scaled back would be drawing the wrong lesson from the right evidence.

The first and most important point is that the counterfactual to imperfect screening is not perfect screening but no screening, and no screening is considerably worse. An organization that assesses its suppliers catches the ones with no policies, no certifications, visible financial distress, and adverse media histories, and those suppliers are meaningfully more likely to cause problems than the ones that pass. The cases in this article are notable precisely because they are cases where assessment failed, and cases where assessment worked do not generate headlines or academic papers. The visible record is systematically biased toward failures, and a reader should discount accordingly.

The second point is that the divergence between raters, while real, does not mean the ratings are noise. An average pairwise correlation around zero point six is a substantial correlation. It means the providers agree considerably more than chance and considerably less than the presentation implies, which is a different claim from saying the measure is meaningless. Ratings do carry signal, particularly at the extremes: suppliers that score badly across multiple providers are identified consistently, and that consistent identification is useful. The problem arises in the middle of the distribution, where the divergence is largest and where threshold decisions are most often made.

The third point is that audits, whatever their limitations, have driven material improvements in conditions where they have been applied persistently and combined with other mechanisms. Factory safety in the garment sector improved substantially in the decade after the 2013 disaster, through binding agreements with inspection regimes that had remedy and enforcement attached, which is different from and better than the audit regime that preceded it. That improvement demonstrates both that the earlier approach was inadequate and that assessment regimes can be redesigned to work better, which is the constructive conclusion this article is aiming at rather than a case for abandonment.

A fourth point deserves emphasis because it constrains what any assessment can achieve. Some failures are not predictable by any method available to an outside party, and holding assessment to a standard of predicting them is holding it to an impossible standard. A supplier concealing fraud, a management team making a catastrophic decision after the assessment, a sudden regulatory change, and a physical disaster are all outside what any scoring model could capture. The reasonable expectation of an assessment programme is that it identifies suppliers with visible problems, directs attention proportionately, and maintains enough contact that emerging problems surface earlier than they otherwise would. Judged by that standard, well-run programmes perform respectably, and the criticism in this article is directed at how their outputs are interpreted rather than at whether they are worth running.

A fifth point in fairness concerns the direction of travel, which has been broadly positive in ways this article's framing can obscure. Assessment methodologies have improved materially over the past decade: providers have expanded their data sources beyond self-disclosure, incorporated incident and

litigation records, and become more transparent about methodology. Regulatory due-diligence regimes in several jurisdictions have shifted obligations from disclosure toward substantive inquiry, which changes what buyers must actually do rather than what they must report. And the academic literature examined in this article exists at all because researchers gained access to large audit datasets, which represents a level of openness the field did not previously permit.

None of that resolves the structural constraint, which is that documented inputs cannot capture undocumented conduct, and this article's argument stands. But a reader should not conclude that supplier assessment is static or uniformly deficient. The instruments are improving, the research base is growing, and the regulatory environment is pushing in a useful direction. The gap this article identifies is between what the instruments can do and how their outputs are read, and that gap is closable by buyers without waiting for any provider to improve anything.

12 Score the attention, not the supplier

The constructive reframing that follows from the diagnosis is a change in what the score is understood to be an answer to. Instead of asking how risky this supplier is, which the instrument cannot reliably answer, ask how much attention this supplier warrants, which it can.

The reframing has immediate practical consequences. Under the first framing, a good score is a conclusion and the supplier moves into the assessed population. Under the second, a good score is an allocation decision and the supplier receives less attention than a poorly-scoring one, which is a statement about the buyer's time rather than about the supplier's soundness. The distinction sounds semantic and it changes behaviour, because an organization that understands its score as an attention allocator does not treat a green rating as evidence and does not report the proportion of green suppliers as a measure of exposure.

The second element is to combine the score with criticality, which many programmes do poorly. A supplier's risk score and its importance to the operation are independent dimensions, and attention should be allocated on both. A low-scoring supplier of a commodity available from twenty other sources warrants little attention, because the consequence of its failure is a procurement inconvenience. A high-scoring sole-source supplier of a critical component warrants substantial attention despite its score, because the consequence of its failure is severe and the score cannot rule the failure out. Programmes that sort by score alone systematically under-attend the second category, which is where the losses concentrate.

The third element is to add signals that scores structurally cannot capture, and the most valuable ones are generated by the buyer's own relationship rather than purchased. Payment behaviour toward the supplier's own suppliers, where observable, is an early indicator of financial stress. Changes in delivery variance, quality trend, responsiveness, and staff turnover at the account level frequently precede formal distress by months. Requests to change payment terms, unusual pressure for early payment, and changes in the people the buyer deals with are signals a rating agency will never see and that the buyer's own operational staff observe routinely. Capturing them requires a mechanism for operational staff to report what they notice, which most organizations lack, and building that mechanism is cheaper and more informative than upgrading a scoring subscription.

A further refinement concerns how divergence between providers should be used, because most organizations that subscribe to more than one source treat disagreement as an inconvenience to be resolved rather than as information. The instinct is to average the ratings, take the more conservative, or pick a primary provider and ignore the rest, all of which discard the signal. A supplier on which two competent providers disagree sharply is a supplier whose assessed profile depends heavily on which observables are examined, and that dependence is itself a finding: it indicates that the readily available evidence supports more than one interpretation.

Treated that way, divergence becomes a useful triage input in its own right. A supplier rated consistently across providers is one where the visible evidence points the same way regardless of method, and the assessment can be relied on to the modest extent any assessment can. A supplier where the ratings

differ by a wide margin warrants a look at why, and the answer is frequently instructive: one provider may be weighting a disclosure the supplier makes well while another is capturing an incident history it does not disclose. Investigating the disagreement takes an analyst an hour and produces more understanding than either rating alone.

13 An assessment protocol, and a scoring rubric

The principles above combine into a protocol an organization can adopt and a rubric a governance body can use to judge whether a supplier risk programme is producing knowledge or paperwork. Figure 9 sets out the discipline.

Figure 9. Using the score as a triage device rather than a verdict: allocate attention, add signals the score cannot see, prefer assurance whose incentives are sound, assess to the tier that fails, and record what the model missed.

The protocol runs as follows. Treat the score as an attention allocator and never as a conclusion, and combine it with criticality so that high-scoring critical suppliers receive attention rather than clearance. Supplement purchased ratings with relationship signals the buyer's own staff observe: payment behaviour, delivery variance, quality trend, responsiveness, and personnel change. Prefer assurance whose incentives are not compromised, which means unannounced over announced, second-party over third-party, and rotated over settled auditor relationships. Map dependency to the depth at which it actually runs for the components that matter, rather than attempting uniform coverage. And maintain a record of failures the programme did and did not anticipate, so that the model can be calibrated against the organization's own experience.

A scoring rubric

The dimensions below distinguish a programme that generates knowledge from one that generates files.

Dimension	Generates knowledge	Generates files
What the score is	An allocator of scarce attention	A verdict on the supplier
Effect of a good score	Less attention, not less doubt	Supplier is cleared and closed
Criticality	Weighed alongside score, independently	Sorting by score alone
Signal sources	Purchased ratings plus own relationship data	Purchased ratings only
Assurance design	Unannounced, rotated, second-party where possible	Announced third-party paid by the audited
Tier coverage	Deep on what matters, shallow elsewhere	Uniform tier one, nothing below
Divergence	Investigated as a signal in itself	Averaged away or unnoticed

Dimension	Generates knowledge	Generates files
Calibration	Own hit and miss record maintained	No record of what was missed

A programme scoring in the left column knows what its instruments can and cannot detect, directs its scarce attention accordingly, and improves over time because it records its own errors. A programme scoring in the right column produces a complete file and a dashboard in which the proportion of green suppliers is reported as though it measured exposure. The rubric does not make suppliers safer. It makes the organization's understanding of its own exposure correspond more closely to the actual position.

14 Keep a record of what you missed

Of everything in the protocol, the element most consistently absent and most valuable over time is the simplest: maintaining a record of which supplier problems the assessment programme anticipated and which it did not.

The practice is uncomplicated. When a supplier problem occurs, whether a failure, a quality collapse, a compliance breach, a financial distress event, or a disruption, record what the supplier's assessment said beforehand, whether the programme had flagged anything, how far in advance any warning appeared, and what signal would have identified the problem earlier had anyone been looking for it. Over a few years this produces something no purchased instrument can supply: an empirical account of how the organization's own assessment performs against its own supply base, which is the only calibration that is actually relevant to its decisions.

The reason this matters more than it appears to is that assessment programmes have no natural feedback loop. A programme that fails to identify a problem generates no signal that it failed, because the problem is attributed to the supplier rather than to the assessment, and the programme continues unchanged. A programme that correctly identifies a problem also generates no signal, because the flagged supplier is managed and the counterfactual is unobservable. Without deliberate record-keeping, the programme accumulates cost and process indefinitely without anyone establishing whether it works, which describes a substantial proportion of supplier risk functions.

The record also enables a question that governance bodies should ask and rarely do. Rather than asking what proportion of suppliers are assessed, which measures activity, ask what proportion of the supplier problems experienced in the past three years were anticipated by the programme, and by how long. That question is answerable from the record and unanswerable without it, and the answer is far more informative about the state of the organization's exposure than any coverage statistic. An organization that finds its programme anticipated a small share of its actual problems has learned something important and actionable, which is that its instruments are pointed at the wrong things, and it can respond by adding the signal types that would have caught what it missed.

The record has a second use that becomes available once a few years of it exist, which is that it supports a proper conversation about how much the assessment programme should cost. Supplier risk functions expand through accretion, adding providers, questionnaires, and process steps in response to individual incidents, and they are almost never asked to justify the marginal addition because nobody can say what any of it catches. An organization with a hit-and-miss record can ask a sharper question: of the problems that occurred, which would have been caught by the instrument we are considering adding, and at what cost per problem anticipated.

That question will sometimes justify spending more and will sometimes reveal that a substantial part of the existing programme catches nothing the organization did not already know. Both answers are useful and neither is available without the record. It is worth emphasising how modest the effort is: the record is a short entry made whenever a supplier problem occurs, capturing what the assessment said and what the earliest available warning would have been. The discipline is in remembering to make the

entry, not in the analysis, and an organization that starts today will have something worth reading within two years.

15 Conclusion: the number is not the knowledge

Supplier risk scoring, environmental and social rating, and social-compliance auditing are the instruments through which most large organizations know their supply base. They produce numbers, the numbers are consumed as knowledge, and the gap between what the numbers contain and what the knowledge would require is the subject of this article. That gap is not a defect in any particular provider's methodology. It follows from the structure of the exercise: a score can only aggregate what is documented and retrievable, and supplier failure originates disproportionately in what is not.

The evidence is unusually good for a claim of this kind. Peer-reviewed analysis of six major rating providers found pairwise correlations averaging around zero point six one against roughly zero point nine for credit ratings, and decomposition showed that the largest contributor was measurement rather than differing values, which forecloses the comfortable explanation. Academic work on supply chain monitoring found third-party audits less effective than buyers' own, with commercial incentive the proposed mechanism. And the case record includes a factory audited under a compliance scheme months before a collapse that killed more than eleven hundred people, a company with nineteen consecutive unqualified audit opinions that entered liquidation with billions in liabilities, and a financing failure that froze ten billion dollars and reached manufacturing operations several steps removed from anything any score examined.

None of this supports abandoning assessment, and this article has set out the serious case that imperfect screening substantially beats none, that ratings carry real signal particularly at the extremes, and that assessment regimes have driven material improvement where they were redesigned with enforcement attached. What it supports is a change in interpretation and in what the programme does with its outputs. Use the score to allocate attention rather than to reach conclusions, and combine it with criticality so that the high-scoring sole-source supplier gets scrutiny rather than clearance. Add the signals the score cannot contain, most of which the organization's own operational staff already observe and nobody currently captures. Prefer unannounced, rotated, and second-party assurance over announced third-party audits paid for by the audited party. Map dependency to depth where the dependency actually matters instead of promising uniform multi-tier coverage that no organization achieves.

And keep the record. An organization that writes down what its programme anticipated and what it missed will, within a few years, know something about its own exposure that no subscription can tell it, and it will be able to answer the question that matters to a board: not how many suppliers have been assessed, but how many of the problems that actually occurred were seen coming. Most organizations cannot answer that today, and the reason is not that the answer is unobtainable but that nobody has been recording it. The number on the dashboard was never the knowledge. What the organization has learned about which of its suppliers actually fail, and what preceded it, is the knowledge, and it accumulates only for those who choose to write it down.

One further observation is worth leaving with a reader who intends to act on this. The changes recommended here are almost entirely free. Reframing the score as an attention allocator costs

nothing. Combining it with criticality requires a column in a spreadsheet the organization already maintains. Capturing relationship signals requires a route for operational staff to report what they already observe. Preferring unannounced and rotated assurance changes how existing audit budget is spent rather than how much. Mapping dependency selectively uses records the organization already holds. And keeping a record of hits and misses takes minutes per incident. None of it requires a new subscription, a new platform, or a larger function, which is worth stating plainly in a field where the default response to an assessment failure is to buy more assessment.

16 Methodology, caveats, and sources

Methodology

- This article draws on peer-reviewed research, regulatory and parliamentary records, insolvency and liquidation filings, and contemporaneous reporting, current to mid-2026. Supply Chain Research is independent and accepts no payment from the rating providers, audit firms, or risk-monitoring vendors discussed.
- Where a figure originates with a party that sells assessment, monitoring, or resilience products, that interest is identified. Primary academic sources are preferred over secondary summaries, which in this field frequently misstate the underlying figures.

Caveats

- The rating-divergence figures describe environmental, social, and governance ratings of large listed companies. They are the best available evidence on inter-rater agreement in non-financial assessment and are not a direct measurement of supplier risk scores, which are less studied.
- No systematic study of supplier risk score predictive accuracy was located. The cases in this article are existence proofs that the failure mode is real, not a sample from which any rate can be inferred, and no rate is claimed.
- Figures on the share of disruption originating below the first tier come from vendors of supply chain resilience and mapping software and are interested sources. Figure 8 is an indicative pattern rather than measured data and is labelled as such.
- Figures 1, 7, and 9 are conceptual illustrations of structure rather than measured data. Figure 3 reproduces a published decomposition; readers should consult the original paper for its full methodology and limitations.
- Audit regimes differ substantially in scope. A social-compliance audit is not a structural or financial assessment, and criticism of what such audits missed should be read as criticism of how buyers interpreted them rather than as an allegation that auditors exceeded or fell short of their defined scope.
- Regulatory and enforcement positions described here change frequently. Organizations with due-diligence obligations should consult current sources and their own counsel rather than relying on this summary.

Sources

[1] Berg, Kolbel and Rigobon. Aggregate Confusion: The Divergence of ESG Ratings, Review of Finance 26(6), 2022.

[2] Short, Toffel and Hugill. Monitoring Global Supply Chains, Strategic Management Journal, 2016.

[3] Ibanez, Palmarozzo, Short and Toffel. Change Agents: Second- and Third-Party Auditor Quality, Harvard Business School working paper, 2025.

[4] UK Parliament. Carillion: Joint report of the Business and Work and Pensions Committees.

[5] UK National Audit Office. Investigation into the government's handling of the collapse of Carillion.

[6] Financial Reporting Council. Sanctions in respect of the audit of Carillion plc.

[7] European Center for Constitutional and Human Rights. The Rana Plaza collapse and the auditing of supplier factories.

[8] UK Parliament, Treasury Committee. Lessons from Greensill Capital.

Additional context drawn from insolvency filings and contemporaneous reporting on the 2024 bankruptcy of a European battery manufacturer, from published material on import enforcement and supply chain due-diligence regimes, and from vendor-published research on multi-tier disruption, which is identified as an interested source wherever used. This article is analysis, not legal, compliance, or procurement advice, and its conclusions should be validated against your own circumstances and obligations before any decision.

Supply Chain Research is an independent, vendor-neutral research platform for supply chain and technology leaders. We accept no payment from the rating providers, audit firms, or risk-monitoring vendors discussed. This article is analysis, not legal, compliance, or procurement advice, and its conclusions should be validated against your own circumstances before any decision.