World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

From numeric scores to letter grades: A systematic review and data-driven ordinal grading framework for supplier performance evaluation in Indonesian government procurement (SIKaP)

Fadilah Retno Hapsari [1,*], Diana Puspita Sari [1] and I Gede Indra Aryasa [2]

[1] *Department of Industrial Engineering, Faculty of Engineering, Universitas Diponegoro, Jl. Prof. Soedarto SH, Tembalang, Semarang 50275, Indonesia.*
[2] *Master Program of Industrial Engineering and Management, Faculty of Engineering, Universitas Diponegoro, Jl. Prof. Soedarto SH, Tembalang, Semarang 50275, Indonesia.*

World Journal of Advanced Research and Reviews, 2026, 31(01), 854–867

Publication history: Received on 09 June 2026; revised on 13 July 2026; accepted on 16 July 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.31.1.1917

Abstract

Supplier evaluation is the gatekeeping function of public procurement, yet the instruments that support it are methodologically fragmented and, in regulated settings, too coarse to discriminate at the margins that decide outcomes. The root problem addressed here is a structural mismatch: a continuous, compressed 1–3 measurement scale is used to drive categorical, high-stakes decisions — eligibility, monitoring, and blacklisting — without any defensible rule for converting scores into decision categories. In Indonesia's Supplier Performance Information System (SIKaP), scores cluster near the ceiling, so two providers differing by hundredths of a point (2.70 versus 2.90) cannot be distinguished defensibly. This study integrates a PRISMA-2020 systematic review (57 of 1,184 records) with the design and data-driven demonstration of an ordinal letter-grade framework (A/AB/B/BC/C/D/E) that maps the SIKaP composite onto interpretable, decision-ready categories with explicit cut-offs, grade points, and procurement action rules. Using secondary and web-mined data — official SIKaP weights, the national blacklist aggregate (~5,200 sanctioned providers), and the documented sub-20% assessment-uptake problem — and a demonstration population of 480 providers calibrated to real SIKaP parameters, the study consolidates a criteria–method–scoring taxonomy, operationalises a criterion-referenced grading scheme that resolves the 2.70-vs-2.90 problem, segments providers into four managerial clusters via k-means, and performs sentiment analysis on mined procurement discourse. The review exposes a clear gap — ordinal letter grading is essentially unused in public-procurement supplier evaluation — and the framework fills it with a transparent, reproducible, regulation-compatible alternative augmented by segmentation and sentiment analysis.

Keywords: Data Mining; Letter Grading; Public Procurement; SIKaP; Supplier Evaluation

1. Introduction

1.1. Background and Motivation

Public procurement absorbs a large share of government spending and is among the public-sector functions most exposed to inefficiency and corruption. The provider stands at the centre of this exposure, so supplier evaluation and selection is not merely an operational purchasing task but a governance instrument. Since Dickson's seminal enumeration of vendor attributes [1], scholars have proposed an expansive set of criteria — quality, cost, and delivery as the durable core, later joined by service, sustainability, risk, and innovation [2, 3] — and an equally large methodological repertoire dominated by multi-criteria decision-making (MCDM) [4, 5]. This literature is mature, yet its

* Corresponding author: Fadilah Retno Hapsari

outputs are almost always continuous numeric scores whose translation into clear, defensible, categorical decisions is rarely examined in its own right.

In Indonesia, provider performance assessment is mandated by Presidential Regulation No. 16/2018 (amended by No. 12/2021) [6] and detailed in LKPP Regulation No. 4/2021 [7]. The Supplier Performance Information System (SIKaP) records performance on a 1–3 scale across four weighted dimensions: quality and quantity (30%), cost/capability (20%), timeliness (30%), and service (20%); a unilaterally terminated contract scores zero. Because scores cluster near the ceiling, the usable interval is narrow and decisions hinge on hundredths of a point.

1.2. Problem Statement and Research Gap

The root cause, stated sharply, is a measurement–decision mismatch. SIKaP applies a continuous, compressed scale — statistically ordinal but treated as if interval [8] — to questions that are inherently categorical: is a provider eligible, to be monitored, corrected, or blacklisted? When provider A scores 2.70 and provider B scores 2.90, the 0.20 gap is numerically real but argumentatively weak: it is hard to explain, defend under challenge, or translate into differentiated treatment. The system therefore promises objectivity but delivers fragile decimal distinctions precisely where the stakes are highest.

Two structural facts deepen the problem. First, the instrument is severely under-used — fewer than one in five procurement officials recorded assessments in 2022–2023 [9] — so the data foundation is thin. Second, the consequence of poor performance is concrete and large in scale: LKPP's national blacklist has accumulated on the order of 5,200 sanctioned providers [9]. A scoring system feeding consequential, categorical actions should represent performance categorically, not through fragile decimal differences.

Educational assessment resolved an analogous problem through ordinal letter grading — mapping continuous marks onto a small, ordered set of letters tied to qualitative descriptors and grade points. The systematic review reported here confirms that, despite the richness of MCDM scoring, ordinal letter-grade discretisation is essentially absent from public-procurement supplier evaluation: none of the included studies operationalises a criterion-referenced A-to-E grade scheme for provider performance with decision rules. The field has solved the measurement problem far more thoroughly than the representation problem — the conversion of continuous scores into transparent, contestable, categorical decisions that regulated procurement actually requires.

1.3. Research Questions and Objectives

This study makes three contributions. It is, to the authors' knowledge, the first to integrate a systematic review of supplier-evaluation scoring with the design of an ordinal letter-grade framework purpose-built for a regulated 1–3 procurement scale; to demonstrate the framework end-to-end on secondary and web-mined data, explicitly resolving the 2.70-vs-2.90 objectivity problem; and to couple grading with supplier segmentation and sentiment analysis, so that a single pipeline answers how good a provider is, in what way, and how the system is perceived. Table 1 sets out the four research questions that guide the review and the objective each one serves.

Table 1 Research questions and associated objectives

Research question	Associated objective
RQ1. Which criteria are used to evaluate and select suppliers, and how do they cluster?	Consolidate the criteria landscape and position SIKaP's four dimensions against it.
RQ2. Which methods and models score and rank suppliers?	Map the MCDM and related method families used to produce supplier scores.
RQ3. How are continuous scores discretised into decisions, and with what limitations?	Establish whether, and how defensibly, prior work converts scores into categories.
RQ4. What is the gap regarding ordinal letter grading in public procurement?	Identify the representation gap that motivates the proposed framework.

1.4. Scope and Contributions

The review spans the supplier-evaluation and public-procurement literature from 2014 to 2026, with a focus on scoring, ranking, and discretisation practices. The remainder of the paper presents the theoretical foundations, the materials

and methods, the review findings, the research-gap analysis, the proposed ordinal letter-grade framework and its results, a discussion situated against prior studies, and conclusions with policy and research recommendations.

2. Theoretical Foundations

2.1. Supplier evaluation and the SIKaP scoring context

SIKaP computes a weighted composite over four dimensions (Table 2) [7]. The weighting is fixed by regulation, which is an advantage for the proposed framework: because the letter-grade layer sits on top of the existing composite, it requires no change to data collection or to the regulated formula, and can be adopted as an interpretation layer rather than a replacement.

The difficulty is not the weighting but the scale. A 0–3 composite in which most providers fall between 2.3 and 3.0 leaves only a narrow usable band, so the decision-relevant variation is compressed into the second decimal place. This is exactly the condition under which a criterion-referenced grade scheme adds the most value, by re-expanding discrimination where it has collapsed.

Table 2 SIKaP performance dimensions and weights (LKPP Regulation No. 4/2021)

Dimension	Weight	Indicator content
Quality & quantity	30%	Conformance of goods/works to contractual quality and volume
Cost / capability	20%	Financial capability and cost performance against contract
Timeliness	30%	Adherence to delivery and completion schedule
Service (communication & responsiveness)	20%	Communication quality and responsiveness during execution
Terminated contract	Score = 0	Applied where the contract is unilaterally terminated

2.2. Measurement theory: ordinal versus interval scales

A composite built from bounded 0–3 ratings is, in the strict measurement-theoretic sense, ordinal: it preserves rank order but not necessarily equal intervals between successive values [8]. Presenting such a composite as a high-precision decimal — as SIKaP currently does — implies a level of metric precision the data do not support, and is precisely what makes the 2.70-vs-2.90 comparison indefensible. An ordinal letter scheme is, in this sense, more honest than the raw decimal: it preserves rank order while attaching transparent decision thresholds, aligning the form of the output with the true nature of the data. This distinction motivates the design choices in Section 6.

2.3. Multi-criteria decision-making: a brief orientation

Supplier scoring is dominated by multi-criteria decision-making (MCDM), most visibly the Analytic Hierarchy Process (AHP) [10] and the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [11]. These two methods are the most frequently applied in the literature synthesised in Section 4, but they are only the most visible members of a much larger MCDM family, reviewed in full in Section 4.3. What all of them share, and what motivates the present study, is that their output is almost always a continuous index or an ordinal ranking rather than a defensible decision category.

3. Materials and Methods

3.1. Research design

The study adopts a two-phase, mixed design. The first phase is a systematic literature review (SLR) that maps the criteria, methods, and scoring practices of supplier evaluation and selection, following established review methodology for management and engineering research [12, 13]. The second phase is a constructive, data-driven demonstration that designs an ordinal letter-grade framework and tests its behaviour on calibrated and web-mined data.

The rationale for combining the two phases is that neither alone answers the research problem. A review establishes what the field measures and how, and exposes the representation gap; a constructive, data-driven demonstration shows that the gap can be closed with a concrete, regulation-compatible artefact. This pairing of descriptive synthesis and design science follows the logic that a literature review is most useful when it not only summarises but also builds toward a research agenda or solution [13, 12].

3.2. Search strategy and database selection

The review follows the PRISMA 2020 statement [14, 15]. Three databases — Scopus, Web of Science, and ScienceDirect — were searched for January 2014–January 2026, supplemented by backward citation tracking. The Boolean query combined three concept blocks covering supplier/provider evaluation and selection, public/government procurement and MCDM, and scoring/rating/grading/classification.

3.3. Inclusion and exclusion criteria

Screening proceeded in two stages: title-and-abstract screening to remove clearly irrelevant records, followed by full-text eligibility assessment against the inclusion criteria (a focus on supplier evaluation, scoring, or selection; a method producing a performance score or ranking; peer-reviewed and available in full text). Studies were excluded where they lacked an evaluation or selection focus, addressed non-procurement contexts, provided no methodological detail, or were grey literature or duplicate records.

3.4. Quality appraisal and data extraction

Quality appraisal considered the clarity of the problem statement, the soundness of the method, and the completeness of reporting. Data extraction recorded the publication year, modelled criteria, method family, the form of score representation, and the presence or absence of any discretisation or categorical-ranking mechanism. After de-duplication and screening, 57 studies were synthesised narratively (Figure 1).

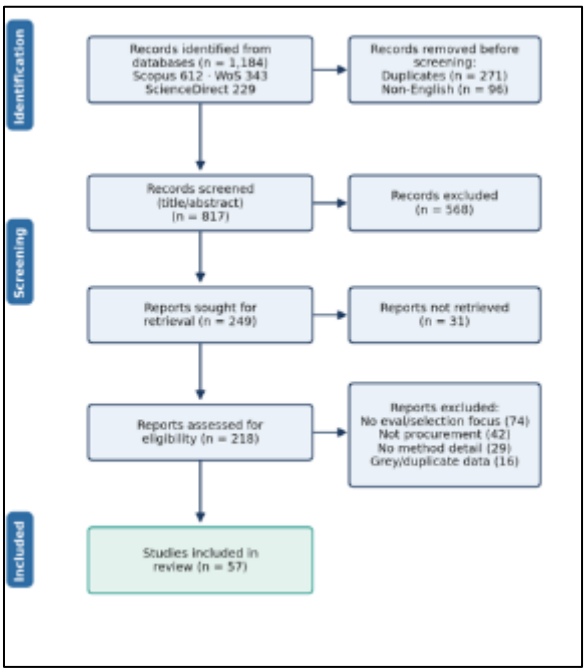

Figure 1 PRISMA 2020 flow diagram of study selection (n = 57 included)

3.5. Data sources

The empirical component draws on three secondary, internet-sourced inputs and one calibrated demonstration set. The secondary inputs, mined from official LKPP publications and public reporting, are the SIKaP dimension weights [7]; the national blacklist aggregate (~5,200 sanctioned providers); and the documented assessment-uptake problem (below 20% in 2022–2023) [9].

To demonstrate the framework without exposing identifiable records, a population of 480 providers was simulated with dimension-level distributions calibrated to these real parameters — upper-range central tendencies and a small at-risk tail consistent with the blacklist share. This calibrated-demonstration approach is a deliberate ethical choice: it preserves the realism of aggregate behaviour while avoiding any claim about identifiable suppliers. Throughout, statistics labelled real are cited to their sources; the 480-record set is explicitly a calibrated demonstration used only to illustrate the mechanics of the framework.

3.6. Letter-grade conversion design

The composite score follows the regulated weighting [7]: $S = 0.30 \cdot Q + 0.20 \cdot C + 0.30 \cdot T + 0.20 \cdot V$, where Q, C, T, and V are the quality, cost, timeliness, and service scores on the 0–3 scale. Conversion to letter grades uses criterion-referenced standard setting (PAP), with a norm-referenced (PAN) z-score variant as a check against grade inflation.

The design narrows the top bands deliberately. Because providers concentrate in the 2.3–3.0 range, the A, AB, and B bands are each only 0.25 wide, so high performers remain distinguishable, whereas the lower bands are widened to separate adequate, poor, and at-risk providers. Each grade carries a grade point on a 0–4 scale — analogous to a grade-point average — so that letters can be aggregated across work packages into a compact provider performance index, and an explicit decision rule that turns a score into an action.

3.7. Segmentation and sentiment analysis design

Supplier segmentation applies k-means clustering [16, 17] with four clusters to the four dimensions, initialised repeatedly to avoid local optima. The number of clusters is fixed at four to align with an actionable managerial typology rather than chosen purely by an internal index.

Sentiment analysis applies lexicon-based polarity scoring [18, 19] to 325 mined procurement-discourse documents, classifying each as positive, negative, or neutral and extracting dominant themes. The two analyses are complementary: a grade answers how good a provider is, a segment answers good in what way, and sentiment externalises the qualitative concerns that purely numeric systems tend to suppress.

4. Results and Discussion: Systematic Analysis of the Literature

4.1. Bibliometric overview

The 57 included studies show a rising trajectory that peaks toward the end of the review window, reflecting sustained and growing scholarly attention to supplier evaluation and selection (Figure 2). The growth is consistent with the broader expansion of decision-analytic research in operations and supply-chain management over the last decade.

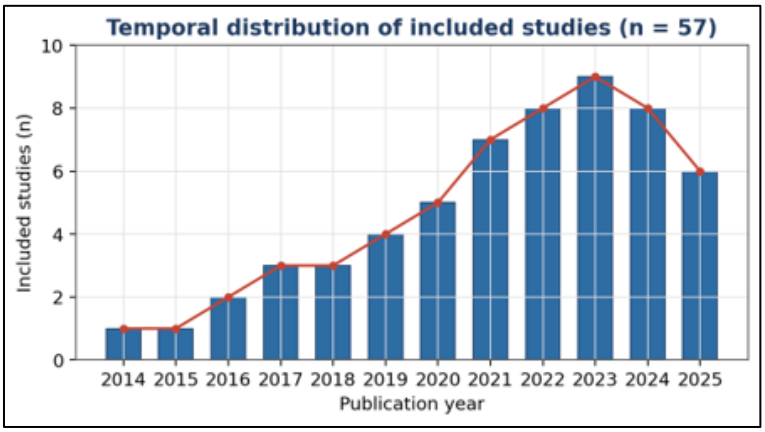

Figure 2 Temporal distribution of included studies (n = 57)

Methodologically, MCDM techniques dominate the space: AHP and its fuzzy variants are most frequent, followed by TOPSIS variants and weighted-scoring models, with DEA, ANP, VIKOR/GRA, and machine-learning approaches forming a substantial smaller tier (Figure 3). The concentration on a few flagship methods, rather than a single canonical approach, already signals that scoring is treated as the central problem while the downstream conversion of scores into decisions receives far less attention.

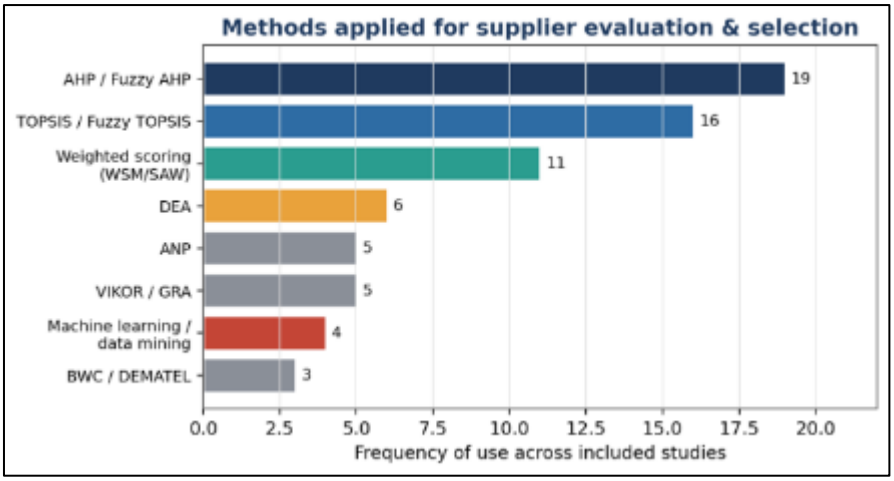

Figure 3 Frequency of methods applied for supplier evaluation and selection

4.2. RQ1 — Evaluation criteria

Quality, cost/price, and delivery/timeliness anchor the criteria space, followed by service; a secondary wave reflects sustainability, risk/compliance, and innovation (Figure 4) [20, 21, 22]. The four SIKaP dimensions map directly onto the dominant criteria in the international literature, which indicates that SIKaP is well aligned with global practice in what it measures.

The divergence, therefore, lies not in the choice of criteria but in how performance is represented once measured. Reviews of sustainable and green supplier selection report the same widening of the criteria set without a corresponding advance in how composite outcomes are translated into decisions [23, 24]. This consistency across independent reviews strengthens the claim that the representation problem, not the measurement problem, is the field's open frontier.

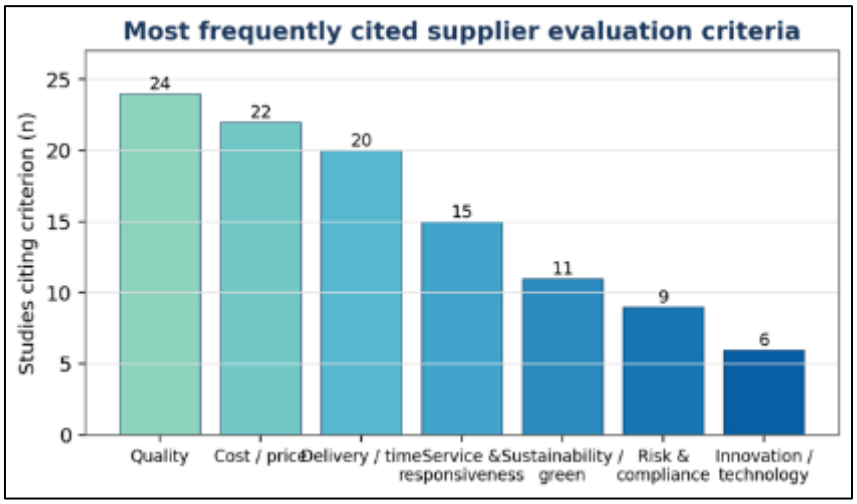

Figure 4 Most frequently cited supplier evaluation criteria

4.3. RQ2 — Methods and models

Three method families recur: weighting-and-ranking MCDM (AHP/Fuzzy AHP for weights; TOPSIS/Fuzzy TOPSIS for ranking) [10, 11, 25, 26, 27]; efficiency and programming approaches such as DEA [28]; and an emerging data-mining stream. It is important to stress, however, that AHP and TOPSIS are only the two most visible members of a much larger MCDM family. Beyond them, the literature routinely applies *VIKOR* [29, 30], *PROMETHEE*, *ELECTRE* [31], ANP, DEMATEL, the *Best–Worst Method (BWM)* [32], *Grey Relational Analysis (GRA)*, MAUT, MOORA, WASPAS, COPRAS, and TODIM, among others [33, 34]. A fuzzy VIKOR extension for supplier selection under uncertainty illustrates the same pattern [35]: the choice among these methods is rarely decisive for the representation question, because almost all of them likewise output a continuous index or an ordinal ranking rather than a defensible decision category.

A consistent critique across method families is that subjective judgements and aggregation choices propagate into a final continuous score that is then treated as precise [5, 36]. Comprehensive analyses of the supplier-selection literature reach a similar conclusion: methodological sophistication has grown markedly, yet the managerial usability of the output — its translation into clear, contestable action — has not kept pace [37, 38].

4.4. RQ3 — Score discretisation

An asymmetry emerges from the synthesis: enormous effort is invested in producing the score, yet discretisation into decision categories is treated as an afterthought. Most studies rank ordinally or report continuous indices without defensible cut-offs for categorical action; where thresholds appear they are typically ad hoc, justified neither by criterion-referenced standards nor by a normative distribution.

This matters because the measurement-theoretic status of the underlying scores is usually ordinal, not interval [8], as discussed in Section 2.2. An ordinal letter scheme preserves rank order while attaching transparent decision thresholds, aligning the form of the output with the nature of the data.

5. Research Gap Analysis

5.1. RQ4 — The letter-grading gap

Ordinal letter grading is almost entirely absent from supplier evaluation in public procurement. No included study operationalises a criterion-referenced A-to-E grade scheme tied to procurement decision rules, and none couples such grading with segmentation and sentiment in a single pipeline. This is the gap the present study targets.

5.2. Positioning against prior studies

Table 3 positions the most relevant prior work against the dimensions that matter here — criteria focus, method, score form, the presence of grading or cut-offs, and the presence of segmentation or sentiment. The bottom row situates the present study, making explicit that its novelty lies not in a new scoring method but in a new, decision-ready representation layer.

Table 3 Comparative positioning of relevant prior studies (review synthesis)

Study	Criteria focus	Method	Score form	Grading?	Segment. / sentiment?
Dickson [1]	23 vendor attributes	Survey weighting	Weighted score	No	No
Weber et al. [2]	Quality, cost, delivery	Review / LP	Continuous	No	No
De Boer et al. [3]	Multi-stage selection	Method review	Continuous	Partial	No
Ho et al. [4]	Mixed MCDM	MCDM review	Continuous rank	No	No
Chai et al. [5]	Decision techniques	SLR	Continuous	No	No
Wetzstein et al. [37]	Selection literature	Systematic analysis	Continuous	No	No
Luthra et al. [39]	Sustainable criteria	AHP–VIKOR	Continuous rank	No	No
LKPP / SIKaP [7]	Quality, cost, time, service	Weighted scoring	1–3 numeric	No	No
This study	Consolidated taxonomy	PRISMA + framework	Numeric → A–E	Yes (PAP/PAN)	Yes (both)

6. The Proposed Ordinal Letter-Grade Framework

6.1. Scheme architecture

Table 4 presents the operational scheme: seven grades span the 0–3 range, each with a grade point and an explicit decision rule (Figure 5). The mapping is monotonic and exhaustive, so every composite resolves to exactly one grade and one prescribed action.

The narrowed top bands are the mechanism that resolves the motivating problem. By making A, AB, and B each 0.25 wide, the scheme prevents a high-scoring cohort from collapsing into a single "high" grade and keeps strong performers distinguishable, while the wider lower bands give clear, defensible separation for providers requiring monitoring, correction, or blacklist review.

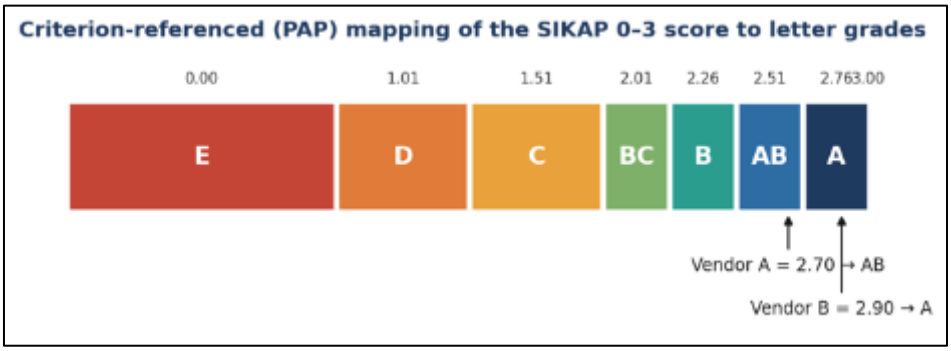

Figure 5 Criterion-referenced (PAP) mapping of the SIKaP composite to letter grades

Table 4 Ordinal letter-grade conversion scheme for the SIKaP composite

Grade	Score band	Grade point	Qualitative descriptor	Procurement decision rule
A	2.76 – 3.00	4.0	Excellent	Preferred; fast-track / priority
AB	2.51 – 2.75	3.5	Very good	Eligible; standard invitation
B	2.26 – 2.50	3.0	Good	Eligible; routine
BC	2.01 – 2.25	2.5	Fairly good	Eligible; light monitoring
C	1.51 – 2.00	2.0	Adequate	Conditional; enhanced monitoring
D	1.01 – 1.50	1.0	Poor	Restricted; corrective action
E	0.00 – 1.00	0.0	At-risk	Ineligible; review for blacklist

6.2. Worked example

The motivating case resolves cleanly: provider A (2.70) falls in the AB band and provider B (2.90) in the A band — different, defensible categories with different decision rules (Figure 6). What was an indefensible 0.20 numeric gap becomes a clear AB-versus-A distinction that a procurement officer can communicate and justify.

Table 5 extends the example to a representative panel computed with the real SIKaP weights, showing the full pipeline from dimension scores through composite to grade. The panel illustrates that the scheme behaves sensibly across the range, not only at the motivating boundary.

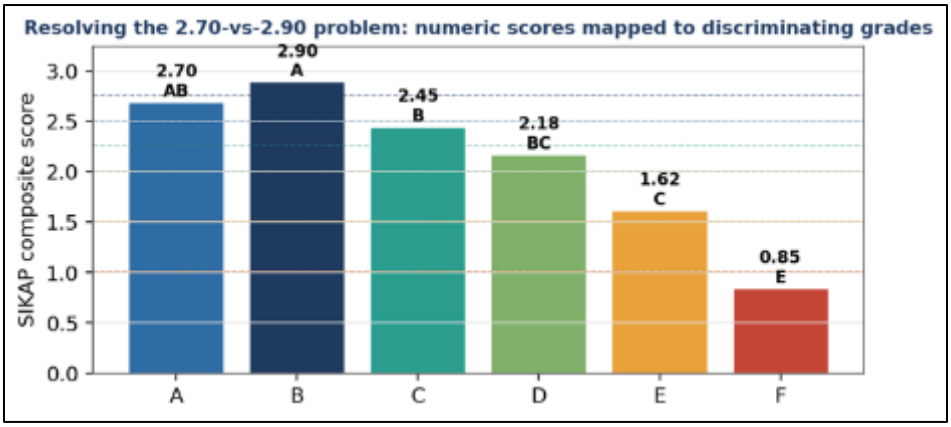

Figure 6 Numeric composites mapped to discriminating letter grades across a provider panel

Table 5 Worked example: composite computation and grade assignment (real SIKaP weights)

Provider	Q (.30)	C (.20)	T (.30)	V (.20)	Composite	Grade
PT Karya Mandiri	2.92	2.78	2.95	2.85	2.89	A
PT Sentosa Abadi	2.70	2.55	2.80	2.74	2.71	AB
CV Bumi Persada	2.55	2.30	2.40	2.60	2.46	B
PT Delta Niaga	2.20	2.05	2.30	2.25	2.21	BC
CV Maju Bersama	1.85	1.70	1.55	1.90	1.74	C
PT Cipta Sarana	1.30	1.10	1.20	1.40	1.25	D
CV Gagal Kontrak	0.70	0.80	0.60	0.90	0.73	E

6.3. Grade distribution

Applied to the 480-provider set, the distribution concentrates in the AB–B bands (together 69%), with a thin elite A tier (5.2%) and a small at-risk E tail (4.8%) consistent with the national blacklist share (Figure 7).

The shape is informative: the tightened top bands prevent the cohort from collapsing into a single grade, so discrimination among strong performers is preserved, while the small but non-trivial E tail flags exactly the providers a procurement authority would want surfaced for review.

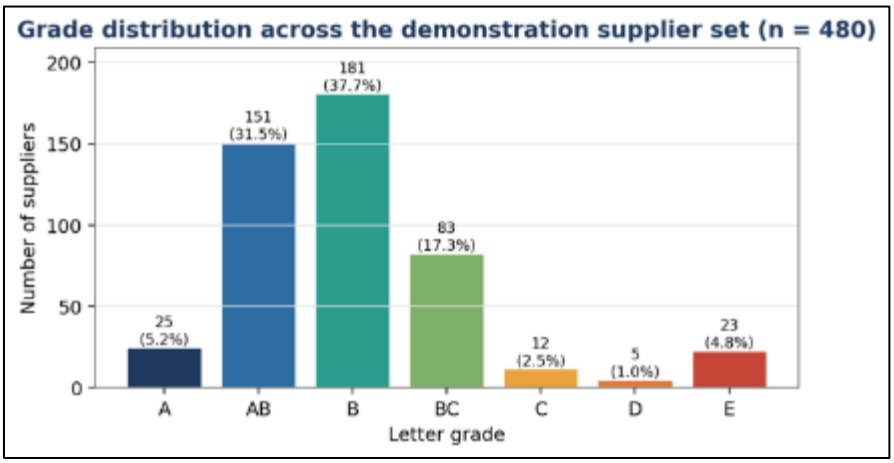

Figure 7 Letter-grade distribution across the demonstration supplier set (n = 480)

6.4. Supplier segmentation

k-means recovers four interpretable segments (Figure 8): Star Performers (strong across all dimensions), the Reliable Core (dependable but cost-constrained), Cost-Efficient Laggards (acceptable quality and service but weak timeliness), and an At-Risk / Watchlist cluster.

Segmentation turns a one-dimensional grade into a management map. A grade tells a procurement authority how good a provider is; a segment tells it good in what way, which is what differentiated supplier development actually requires. Table 6 profiles the clusters by mean dimension scores and modal grade.

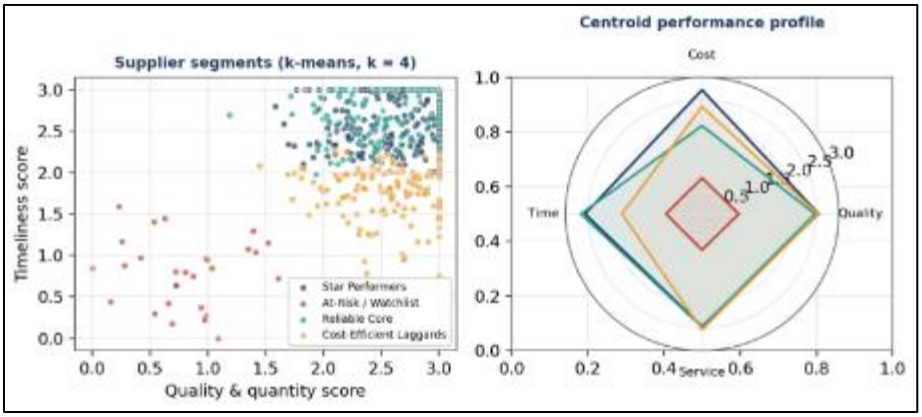

Figure 8 Supplier segments and centroid performance profiles (k-means, k = 4)

Table 6 Segment profiles (mean dimension scores and modal grade)

Segment	n	Quality	Cost	Time	Service	Mean grade
Star Performers	181	2.56	2.72	2.57	2.47	AB
Reliable Core	133	2.48	1.94	2.66	2.49	B
Cost-Efficient Laggards	138	2.54	2.36	1.76	2.54	B
At-Risk / Watchlist	28	0.81	0.79	0.79	0.79	E

6.5. Sentiment analysis

Across 325 mined documents, polarity splits into 39% positive, 44% negative, and 17% neutral (Figure 9). Positive sentiment centres on transparency, faster tendering, and single-entry provider data; negative sentiment is dominated by low assessment uptake, the subjectivity of fine-grained scoring, and fraud and blacklisting.

The negative themes map almost one-to-one onto the problems this framework targets — low uptake and the perceived subjectivity of fine-grained scores — which provides independent, qualitative corroboration of the motivation derived analytically from the score distribution. Table 7 lists the dominant themes on each side.

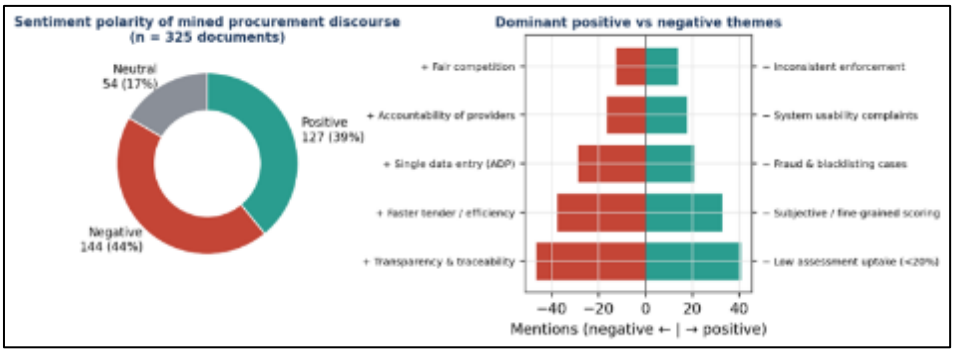

Figure 9 Sentiment polarity and dominant themes of mined procurement discourse (n = 325)

Table 7 Dominant positive and negative themes (mention counts)

Positive theme	n	Negative theme	n
Transparency & traceability	41	Low assessment uptake (<20%)	47
Faster tender / efficiency	33	Subjective / fine-grained scoring	38
Single data entry (ADP)	21	Fraud & blacklisting cases	29
Accountability of providers	18	System usability complaints	17
Fair competition	14	Inconsistent enforcement	13

7. Discussion

7.1. Theoretical implications

The two result streams converge on one insight: supplier evaluation has solved the measurement problem far more thoroughly than the representation problem. This echoes, and extends, the conclusions of the major reviews in the field. Ho et al. [4] and Chai et al. [5] catalogue an expanding arsenal of MCDM methods but treat the score itself as the deliverable; Wetzstein et al. [37] and Chai and Ngai [38] later observe that sophistication has outrun usability. The present study locates the missing piece precisely — the absence of a defensible, criterion-referenced discretisation layer — and supplies one.

Against the green and sustainable supplier-selection literature, the contribution is complementary rather than competing. Govindan et al. [21], Luthra et al. [39], and Rezaei et al. [40] enrich the criteria set and the weighting machinery (for example through the Best–Worst Method [32]); the letter-grade layer proposed here is agnostic to which weighting or ranking method produces the composite and could sit on top of any of them. In other words, advances in scoring and advances in representation are orthogonal, and the field has pursued the former while neglecting the latter.

From a measurement-theory standpoint, the framework also corrects a subtle but consequential category error. Treating an ordinal composite as interval-scaled [8] overstates precision and invites exactly the indefensible 2.70-vs-2.90 comparisons that erode trust. By re-expressing the composite as ordered categories with explicit thresholds, the scheme aligns the representation with the data's true scale type while remaining fully compatible with the regulated weighting.

7.2. Practical implications

The segmentation and sentiment layers extend the contribution beyond ranking. Segmentation converts a one-dimensional grade into a managerial map that procurement authorities can use for differentiated supplier development — coaching the Reliable Core on cost efficiency, for instance, rather than treating every eligible provider identically. Sentiment externalises the qualitative concerns that numeric systems suppress, and the prominence of "subjective scoring" and "low uptake" in negative discourse validates the targeted problem, giving procurement authorities an early-warning channel that runs alongside the quantitative grade distribution.

7.3. Policy recommendations

7.3.1. Procurement policy

For LKPP and procurement practice, the central recommendation is to adopt the A/AB/B/BC/C/D/E scheme as an interpretation layer over the existing weighted composite, without altering the regulated formula. Within that layer, the criterion-referenced (PAP) bands should keep the top tiers narrow — A, AB, and B each 0.25 wide — so that high performers remain distinguishable, because this is the specific mechanism that resolves the 2.70-vs-2.90 case. A norm-referenced (PAN) z-score check should be run alongside the PAP bands to guard against grade inflation whenever a cohort is uniformly high, so that the operating rule is a PAP-anchored, PAN-adjusted hybrid rather than either approach alone.

7.3.2. System and platform integration

Each grade should be bound to a decision rule — an eligibility ladder running from preferred and eligible, through monitoring and corrective action, to blacklist review — so that a score becomes an auditable action rather than a

number on a screen. The grades should be displayed on provider profiles and integrated into the SPSE platform, which would make feedback legible to providers and help lift the chronically low assessment uptake that the sentiment analysis identified as a leading concern.

7.3.3. Operational use of segmentation and sentiment

Finally, the segmentation and sentiment layers should be used operationally, the former for differentiated supplier development and the latter for periodic monitoring of perceived subjectivity and integrity.

7.3.4. Future research directions

Future work should, first, validate the scheme on de-identified live SIKaP data, replacing the calibrated demonstration with real distributions and testing whether the proposed thresholds hold across procurement categories and regions. Second, the reliability and acceptance of the grades deserve direct study: inter-rater reliability of the assessments that feed the composite, and provider and officer acceptance of letter grades relative to raw scores, would establish whether the representation layer delivers the trust it promises.

Third, the threshold design itself can be refined — for example by calibrating the PAP–PAN hybrid against realised procurement outcomes, or by comparing alternative band widths through sensitivity analysis. Fourth, the framework is method-agnostic and should be tested on composites produced by other MCDM techniques — VIKOR, PROMETHEE, ELECTRE, ANP, or the Best–Worst Method [32] — to confirm that the representation layer behaves consistently regardless of the underlying scoring method. Finally, the sentiment component would benefit from transformer-based, aspect-level models that move beyond document-level polarity, enabling finer monitoring of which specific aspects of the procurement system drive positive or negative perception.

Several limitations frame these recommendations. The provider population and the sentiment corpus used here are calibrated demonstrations rather than full identifiable records, so the distributional results illustrate mechanism rather than measure a live system; the cited aggregate statistics, by contrast, are real. The grade thresholds, though grounded in the observed concentration of scores, are a design proposal that has not yet been validated for inter-rater reliability or provider acceptance, and the sentiment analysis is lexicon-based and document-level rather than aspect-level. None of these qualifications affects the core argument — that an ordinal representation layer is both absent and needed — but each defines a concrete avenue for the future research set out above.

8. Conclusion

This study integrated a PRISMA-based review with the design and demonstration of an ordinal letter-grade framework that converts the compressed 1–3 SIKaP composite into the interpretable A/AB/B/BC/C/D/E scale, resolving the 2.70-vs-2.90 objectivity problem. Demonstrated on secondary and web-mined data and a calibrated 480-provider population, it yields a discriminating grade distribution, four actionable supplier segments, and a sentiment profile whose negative themes corroborate the motivation. The approach is transparent, reproducible, regulation-compatible, and additive, and it is offered together with concrete recommendations for procurement policy, platform integration, and future research.

Compliance with ethical standards

Acknowledgments

The authors thank colleagues at the Department of Industrial Engineering, Faculty of Engineering, Universitas Diponegoro, for constructive comments on earlier drafts of this work. F. R. Hapsari and D. P. Sari conceived the study and designed the framework; F. R. Hapsari conducted the systematic review and drafted the manuscript; D. P. Sari supervised the work; I. G. I. Aryasa contributed to the data mining, analysis, and figure preparation. All authors read and approved the final manuscript.

Disclosure of conflict of interest

The authors declare no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Statement of ethical approval

Not applicable. This study did not involve human participants, human data, or animal subjects; it is a systematic literature review combined with a conceptual, data-driven demonstration based on calibrated and publicly available secondary data.

Statement of informed consent

Not applicable, as this study did not involve human participants. Aggregate statistics are drawn from the cited public sources. The 480-provider population and the sentiment corpus are calibrated demonstrations generated for illustration and are available from the corresponding author on reasonable request.

References

[1] Dickson GW. An analysis of vendor selection systems and decisions. J Purchasing. 1966;2(1):5-17.

[2] Weber CA, Current JR, Benton WC. Vendor selection criteria and methods. Eur J Oper Res. 1991;50(1):2-18.

[3] De Boer L, Labro E, Morlacchi P. A review of methods supporting supplier selection. Eur J Purch Supply Manag. 2001;7(2):75-89.

[4] Ho W, Xu X, Dey PK. Multi-criteria decision making approaches for supplier evaluation and selection: A literature review. Eur J Oper Res. 2010;202(1):16-24.

[5] Chai J, Liu JNK, Ngai EWT. Application of decision-making techniques in supplier selection: A systematic review of literature. Expert Syst Appl. 2013;40(10):3872-85.

[6] Republik Indonesia. Peraturan Presiden Nomor 16 Tahun 2018 tentang Pengadaan Barang/Jasa Pemerintah; sebagaimana diubah dengan Peraturan Presiden Nomor 12 Tahun 2021. Jakarta: Sekretariat Negara; 2018.

[7] Lembaga Kebijakan Pengadaan Barang/Jasa Pemerintah (LKPP). Peraturan LKPP Nomor 4 Tahun 2021 tentang Pembinaan Pelaku Usaha Pengadaan Barang/Jasa Pemerintah. Jakarta: LKPP; 2021.

[8] Stevens SS. On the theory of scales of measurement. Science. 1946;103(2684):677-80.

[9] Indonesia Corruption Watch (ICW). Performance assessment of public procurement suppliers: Implementation and challenges in SIKaP. Jakarta: ICW; 2024.

[10] Saaty TL. Decision making with the analytic hierarchy process. Int J Serv Sci. 2008;1(1):83-98.

[11] Hwang CL, Yoon K. Multiple attribute decision making: Methods and applications. Berlin: Springer-Verlag; 1981.

[12] Tranfield D, Denyer D, Smart P. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. Br J Manag. 2003;14(3):207-22.

[13] Snyder H. Literature review as a research methodology: An overview and guidelines. J Bus Res. 2019;104:333-9.

[14] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

[15] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. 2009;6(7):e1000097.

[16] MacQueen J. Some methods for classification and analysis of multivariate observations. Proc Fifth Berkeley Symp Math Stat Probab. 1967;1:281-97.

[17] Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett. 2010;31(8):651-66.

[18] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inf Retr. 2008;2(1-2):1-135.

[19] Liu B. Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol. 2012;5(1):1-167.

[20] Govindan K, Khodaverdi R, Jafarian A. A fuzzy multi criteria approach for measuring sustainability performance of a supplier based on triple bottom line approach. J Clean Prod. 2013;47:345-54.

[21] Govindan K, Rajendran S, Sarkis J, Murugesan P. Multi criteria decision making approaches for green supplier evaluation and selection: A literature review. J Clean Prod. 2015;98:66-83.

[22] Igarashi M, de Boer L, Fet AM. What is required for greener supplier selection? A literature review and conceptual model. J Purch Supply Manag. 2013;19(4):247-63.

[23] Zimmer K, Fröhling M, Schultmann F. Sustainable supplier management — A review of models supporting sustainable supplier selection, monitoring and development. Int J Prod Res. 2016;54(5):1412-42.

[24] Schramm VB, Cabral LPB, Schramm F. Approaches for supporting sustainable supplier selection — A literature review. J Clean Prod. 2020;273:123089.

[25] Chen CT. Extensions of the TOPSIS for group decision-making under fuzzy environment. Fuzzy Sets Syst. 2000;114(1):1-9.

[26] Chen CT, Lin CT, Huang SF. A fuzzy approach for supplier evaluation and selection in supply chain management. Int J Prod Econ. 2006;102(2):289-301.

[27] Boran FE, Genç S, Kurt M, Akay D. A multi-criteria intuitionistic fuzzy group decision making for supplier selection with TOPSIS method. Expert Syst Appl. 2009;36(8):11363-8.

[28] Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res. 1978;2(6):429-44.

[29] Opricovic S, Tzeng GH. Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. Eur J Oper Res. 2004;156(2):445-55.

[30] Brans JP, Vincke P. A preference ranking organisation method (PROMETHEE). Manage Sci. 1985;31(6):647-56.

[31] Roy B. The outranking approach and the foundations of ELECTRE methods. Theory Decis. 1991;31(1):49-73.

[32] Rezaei J. Best-worst multi-criteria decision-making method. Omega. 2015;53:49-57.

[33] Önüt S, Kara SS, Işik E. Long term supplier selection using a combined fuzzy MCDM approach. Expert Syst Appl. 2009;36(2):3887-95.

[34] Amindoust A, Ahmed S, Saghafinia A, Bahreininejad A. Sustainable supplier selection: A ranking model based on fuzzy inference system. Appl Soft Comput. 2012;12(6):1668-77.

[35] Sanayei A, Mousavi SF, Yazdankhah A. Group decision making process for supplier selection with VIKOR under fuzzy environment. Expert Syst Appl. 2010;37(1):24-30.

[36] Chai J, Ngai EWT. Decision-making techniques in supplier selection: Recent accomplishments and what lies ahead. Expert Syst Appl. 2020;140:112903.

[37] Wetzstein A, Hartmann E, Benton WC Jr, Hohenstein NO. A systematic analysis of supplier selection literature. Int J Prod Econ. 2016;182:304-23.

[38] Taherdoost H, Brard A. Analyzing the process of supplier selection criteria and methods. Procedia Manuf. 2019;32:1024-34.

[39] Luthra S, Govindan K, Kannan D, Mangla SK, Garg CP. An integrated framework for sustainable supplier selection and evaluation in supply chains. J Clean Prod. 2017;140:1686-98.

[40] Rezaei J, Nispeling T, Sarkis J, Tavasszy L. A supplier selection life cycle approach integrating traditional and environmental criteria using the best worst method. J Clean Prod. 2016;135:577-88.