Operations Research Forum (2026) 7:129
https://doi.org/10.1007/s43069-026-00715-1

RESEARCH

A Framework for SDG Evaluation Integrating 2-Tuple Linguistic Aggregation with Choquet Capacities and Deck-of-Cards Elicitation: Methodology and Decision Support Tool

F.J. Martínez[1] · Silvia Cruz[1] · M. Espinilla[1]

Received: 17 August 2026 / Accepted: 14 October 2026
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2026

Abstract

Evaluating Sustainable Development Goals (SDGs) requires integrating heterogeneous expert judgments across qualitative sustainability dimensions—a task for which traditional numerical frameworks cause information loss and representational bias. We propose a heterogeneous fuzzy linguistic model (HFLM) based on the 2-tuple linguistic representation model and the Computing with Words (CWW) paradigm and present a decision support tool implementing it end-to-end. The framework supports three aggregation levels: expert assessments are consolidated per criterion via 2-tuple weighted average or ordered weighted averaging (OWA) operators; criterion results are combined into a global evaluation via a Choquet integral governed by a 2-additive capacity, capturing non-additive SDG interactions. The capacity is built from fuzzy densities obtained through a structured Deck-of-Cards (Simos–Roy–Figueira) protocol together with pairwise interaction terms from targeted specialist judgement. The framework is first verified computationally against analytical baselines, then subjected to a controlled simulation experiment under a factorial design, and finally illustrated through a fully synthetic case study—a simulated 50-expert multidisciplinary panel evaluating Internet of Things (IoT)-enabled healthcare for ageing populations (SDG 3)—designed to test its mathematical robustness, heterogeneity handling, and discriminative capacity under controlled conditions. The results show that the Choquet aggregation identifies Rural Connectivity as the structural bottleneck of the synthetic scenario and quantifies its amplification, displacing the global score below every weighted-average baseline without requiring the blanket pessimism of an OWA operator, while the tool is designed to make this discriminative capacity accessible to practitioners without specialist knowledge of the formalism.

✉ F.J. Martínez
fjmimbre@ujaen.es

Silvia Cruz
scr00043@red.ujaen.es

M. Espinilla
mestevez@ujaen.es

[1] Computer Science Department, University of Jaén, Campus Las Lagunillas s/n, Jaén 23071, Spain

Published online: 05 November 2026

Keywords Computing with Words · Choquet integral · k-additive capacities · Multi-granular linguistic information · Deck-of-Cards method · Sustainable Development Goals

1 Introduction

The evaluation of complex real-world phenomena involving uncertainty, imprecision, and qualitative judgement has motivated the development of linguistic decision models as a principled alternative to crisp numerical approaches [1]. Within this paradigm, the 2-tuple linguistic representation model [2] has established itself as a particularly robust framework for Computing with Words (CWW), offering exact symbolic computation with no additional numerical loss within the 2-tuple representation—in contrast to earlier ordinal or extension-principle-based approaches, which require rounding at each aggregation step [2]. Extensions of this model to multi-granular and heterogeneous contexts have further expanded its applicability to group decision-making (GDM) problems in which experts operate over linguistic term sets (LTS) of different cardinality or representational domain [3, 4].

In this context, the monitoring and evaluation of progress toward the United Nations Sustainable Development Goals (SDGs) constitute a challenging real-world application domain [5]. The 2030 Agenda, structured around 17 SDGs and 169 associated targets (listed in Online Resource 1, Table S1), requires the systematic integration of heterogeneous expert assessments across dimensions—social, environmental, and economic—that are inherently qualitative and resistant to crisp quantification [6]. Prevailing monitoring mechanisms, based predominantly on official statistical indicators [5], are ill-suited to represent concepts such as institutional resilience or social equity without material loss of semantic content.

Multi-criteria decision-making (MCDM) approaches and fuzzy linguistic models have been proposed to address these representational limitations across a broad range of sustainability-related evaluation contexts [7]. However, a fundamental challenge persists: in practice, expert panels evaluating SDG compliance are inherently heterogeneous, with evaluators operating over different representational domains—numerical scores, interval-valued estimates, and linguistic scales of varying granularity. Enforcing homogeneous representation prior to aggregation, as required by most existing approaches, introduces transformation bias and irreversibly degrades informational richness [4].

Three research lines bear on this problem, each leaving a different part of it unresolved. *Operational SDG assessment frameworks*—the UN composite indicators, the SDG Index and their MCDM-based variants [5, 8, 9]—deliver practitioner-facing outputs, but require homogeneous numerical input and aggregate additively. *Heterogeneous multi-granular linguistic models* [3, 4, 10] resolve the representational problem, but build their aggregation layer on weighted averages and therefore inherit the independence assumption. *Choquet-integral MCDA*, including the k-additive tradition [11, 12] and the Deck-of-Cards capacity identification of Bottero, Barbati and Corrente [13–15], models interaction explicitly and supplies mature elicitation machinery, but is designed for a single decision maker over a common cardinal domain. The closest

prior work [16] combines the first two under a static Choquet integral, yet specifies the fuzzy measure directly rather than through an auditable elicitation protocol, and is not accompanied by an operational implementation.

The methodological gap follows from this diagnosis: *no existing framework carries heterogeneous, multi-granular assessments from a large multidisciplinary panel into a non-additive aggregation whose capacity is elicited through a structured and auditable protocol, and returns the result on the original linguistic scale without rounding loss.* Closing it is not a matter of assembling the three lines side by side, since each imposes a constraint that the others must satisfy; Sect. 2.2 sets out the comparison requirement by requirement. A practical gap compounds the methodological one: existing proposals are typically presented as theoretical frameworks, without operational tools that practitioners can apply without specialist knowledge of the formalism.

This paper addresses these challenges by proposing an SDG assessment framework based on heterogeneous fuzzy linguistic modelling under the 2-tuple paradigm, and by presenting a decision support tool that implements the framework end-to-end. The framework supports three complementary aggregation levels: at the expert level, individual assessments per criterion are consolidated using the 2-tuple weighted average or OWA operators; at the criterion level, panel results are organised and filtered per dimension; and at the global level, criteria are combined via a Choquet integral governed by a 2-additive capacity that captures non-additive SDG interactions. The capacity combines fuzzy densities elicited through a structured Deck-of-Cards procedure [17, 18] with pairwise interaction terms from targeted expert judgement, grounding the construction of the fuzzy measure in established preference-elicitation practice rather than ad hoc weight assignment. The tool is designed to make this entire pipeline directly accessible to non-specialist practitioners. The contribution is therefore not the novelty of the individual components, each of which is established, but the resolution of the gap identified above: to our knowledge no previously reported framework satisfies all five of the requirements compared in Table 1, and none does so within an operational SDG evaluation tool.

The main contributions of this work are as follows:

- **Methodological integration.** A principled operational framework for integrating heterogeneous evaluation domains—numerical, interval-valued, and multi-granular linguistic—within a unified SDG assessment architecture, combining a three-level linguistic aggregation pipeline (expert, criterion, and global) with a Choquet integral governed by a 2-additive capacity whose fuzzy densities are elicited through a structured Deck-of-Cards procedure [17, 18]. Each component—the 2-tuple heterogeneous unification [2, 10], the multi-level aggregation operators [4], the Choquet integral over k-additive capacities [12], and the Deck-of-Cards protocol [17]—is individually established; the contribution lies in their integration within a single end-to-end methodology for SDG evaluation. The integration is not merely additive: the Choquet integral requires criterion values on a common cardinal scale, which the heterogeneous multi-granular setting does not natively provide, while the 2-tuple machinery supplies exactly such a scale through Δ_S^{-1} without the rounding that would otherwise propagate into the capacity-weighted aggregation. Conversely, the Deck-of-Cards protocol yields the ratio-scale densi-

Table 1 Positioning of the proposed framework against the principal families of existing approaches. R1: heterogeneous multi-granular input. R2: non-additive aggregation. R3: structured capacity elicitation. R4: exact linguistic output. R5: operational tool. •, satisfied; ○, partially satisfied or satisfied under restrictive conditions; —, not addressed

Family of approaches	R1	R2	R3	R4	R5
Official and composite SDG monitoring [5, 8, 19]	—	—	—	—	•
Fuzzy and MCDM-based SDG assessment (AHP, TOPSIS variants) [9, 33]	○	—	○	—	—
Heterogeneous multi-granular 2-tuple models [3, 4, 10]	•	—	—	•	—
Choquet and k-additive aggregation in MCDA [11, 12, 31]	—	•	—	—	—
DCM-based capacity identification [13–15]	—	•	•	—	—
Heterogeneous information with static Choquet integral [16]	•	•	—	•	—
Proposed framework (this paper)	•	•	•	•	•

ties that the 2-additive Möbius representation needs, replacing the ad-hoc weight assignment prevalent in operational SDG tools. Each component thus resolves a constraint that the others impose.

- **Prototype implementation.** A decision support tool implementing the complete pipeline, demonstrating that the methodology can be made directly accessible to practitioners: the tool allows configuration of heterogeneous expert panels, assessment entry in native formats, and production of interpretable linguistic SDG evaluations at all three aggregation levels.
- **Computational verification and illustrative application.** The framework is first verified against analytical baselines under degenerate configurations (Sect. 5.1), then subjected to a controlled simulation experiment under a 3×2 factorial design, complemented by scenarios that vary the density profile and the placement of the interacting pair, which together confirm its predicted behaviour under varied interaction structures (Sect. 5.2), and finally illustrated through a fully synthetic case study—with the complete dataset published and all results exactly reproducible from it—comprising a simulated 50-expert panel on IoT-enabled healthcare for ageing populations under SDG 3, demonstrating the discriminative capacity of the Choquet aggregation relative to additive baselines (Sect. 5).

The remainder of this paper is organised as follows. Section 2 surveys existing SDG assessment models and characterises their methodological limitations. Section 3 establishes the mathematical foundations of the heterogeneous fuzzy linguistic approach and the three-level aggregation model. Section 4 presents the proposed operational framework and the decision support tool. Section 5 provides a computational verification of the implementation, a controlled simulation experiment, and an illustrative application under SDG 3. Section 6 provides comparative analysis, Sect. 7 states the limitations and scope of the study, and Sect. 8 draws conclusions and outlines future research.

2 SDGs and the Evaluation Process: Critical Challenges

The monitoring of the 2030 Agenda operates through a multi-level review architecture spanning global, regional, and national scales. At the international level, the primary mechanism is the United Nations Global Indicator Framework (GIAF) [5], which currently encompasses 231 unique indicators designed to track progress toward the 169 SDG targets. Data collection is decentralised, relying on national statistical offices reporting to UN custodian agencies, and feeds high-profile outputs such as the annual *SDG Progress Report* and the *Voluntary National Reviews* presented at the High-Level Political Forum (HLPF) [6]. Complementing these institutional mechanisms, the academic community has developed composite indices—most notably the SDG Index [8]—aggregating diverse indicators into unified cross-country performance scores.

Beyond institutional monitoring, a growing body of computational research has explored data-driven and multi-criteria approaches for SDG evaluation, including composite indicator construction, machine learning analysis, and decision support modelling [19–21]. Despite their utility for macro-level benchmarking, prevailing approaches exhibit structural limitations when confronting the intrinsic complexity, subjectivity, and interdisciplinary nature of sustainability assessment. Three critical gaps motivate the adoption of more flexible computational paradigms.

Epistemic Uncertainty and Over-Reliance on Crisp Data Official and index-based frameworks rely predominantly on quantitative metrics. However, central SDG dimensions—social cohesion, institutional trust, or perceived quality of life—are inherently qualitative and resistant to precise numerical encoding. Representing such constructs through crisp proxies introduces artificial certainty and systematically obscures epistemic uncertainty [22]. Fuzzy and linguistic modelling has been proposed as a principled mitigation strategy [23], yet its integration into operational SDG analytics remains limited.

Information Loss Through Reductionist Aggregation Assessment processes routinely compress multidimensional sustainability performance into binary classifications or linear weighted sums, neglecting non-linear interdependencies and synergies among goals—a structural problem extensively documented in the SDG systems modelling literature [24]. Classical MCDM techniques applied in this context, including TOPSIS and AHP variants, further risk degrading semantic richness during normalisation and aggregation [9].

Heterogeneous Knowledge Representation Comprehensive SDG evaluation requires integrating assessments from multidisciplinary experts who naturally express judgments in heterogeneous formats: precise numerical values, interval estimates, and linguistic terms of varying granularity. Existing fuzzy MCDM frameworks typically impose homogeneous input structures or fixed linguistic granularity [25], resulting in semantic information loss at the representational boundary. Recent advances in multi-granular linguistic decision-making partially address this limitation [26, 27], but practical aggregation mechanisms capable of preserving expressive diversity across

representational domains remain underdeveloped. This gap constitutes the primary technical motivation of the present work.

A complementary strand of the multi-criteria decision aiding (MCDA) literature has focused on the *elicitation* of the preference parameters that govern aggregation, most prominently through the Deck-of-Cards Method (DCM) originating in Simos' procedure and its revision by Figueira and Roy [17]. In the DCM, the decision maker rank-orders a set of cards representing the objects to be evaluated and inserts blank cards between consecutive positions to express preference intensity, from which ratio-scale weights are derived. The method has been extended with pairwise comparison tables and inconsistency management [18], adapted to construct the capacities of Choquet integral models in real-world applications [13], combined with ordinal regression to infer complete value functions—including Choquet-based ones—from holistic preference statements [14], and generalised to hierarchical, robust, and stochastic settings [15]. These developments provide mature machinery for parameter elicitation; however, they address the preferences of a single decision maker (or a small homogeneous group) expressed over a common representational domain, and do not address the unification of heterogeneous, multi-granular assessments from large multidisciplinary panels—the representational problem targeted here. The two strands are therefore complementary: in this work, the DCM is adopted as the elicitation front-end for the fuzzy densities of the 2-additive capacity (Sect. 5.7), while the heterogeneous 2-tuple machinery handles the representational unification that the DCM literature leaves out of scope.

2.1 Recent Developments in Uncertainty-Aware MCDM

A substantial recent literature addresses uncertainty in multi-criteria decision-making through richer representations of individual judgement. The orthopair family generalises intuitionistic fuzzy sets by progressively relaxing the constraint linking membership and non-membership degrees—through Pythagorean, q-rung orthopair and Fermatean formulations, and more recently fractional orthopair sets—so that an expert can express hesitation and reluctance over a wider admissible region than the classical formulation permits. These representations are commonly paired with distance-based ranking procedures, above all TOPSIS, and with entropy-derived criterion weights: recent instances combine fractional orthopair sets, entropy weighting and TOPSIS to rank green supply-chain strategies under complex uncertainty [28], and pair TOPSIS with a sine-based aggregation operator for the robust evaluation of data-science models [29]. The family has also been applied to sustainability assessment directly, in a fractional orthopair framework for urban water resource management [30]. Closely related work brings the Choquet integral into this setting: Lourenzutti and Krohling develop Choquet-based TOPSIS and TODIM procedures for dynamic and heterogeneous decision problems [31, 32], and distance-based group decision models under interval and hesitant representations continue to be refined [27, 33].

This literature and the present framework address two different axes of uncertainty. The orthopair family deepens the representation of uncertainty *within* a single expert's judgement on a single criterion: how confidently a membership degree is asserted. The problem addressed here is uncertainty *across* formats and experts, where assessments

arrive as numbers, intervals and linguistic terms drawn from scales of different cardinality and must be combined without forcing a common format first. The comparison is concrete rather than abstract in the case of [30], which addresses sustainability assessment as we do: it represents the uncertainty of each individual judgement with considerable precision, but requires all assessments to arrive in a single representational format, whereas our problem begins one step earlier. The two axes are orthogonal rather than competing, and an orthopair representation could in principle be substituted at the expert level of the present pipeline without altering the criterion-level or global layers.

A second distinction concerns the decision problem itself. TOPSIS and its extensions are *relative* ranking procedures: they order a set of alternatives by distance to ideal and anti-ideal points, and the resulting closeness coefficient is meaningful only with respect to the particular set evaluated. This holds even when the task is framed as evaluation rather than selection—as in the robust model evaluation of [29], where the output remains a coefficient defined against the comparison set. SDG assessment as posed here is an *absolute* evaluation problem—the object is to place one evaluated unit on an interpretable scale, not to rank it against competitors—and the required output is a linguistic term that a practitioner can report. This difference in the decision problem, rather than any claim of superiority in ranking accuracy, is what separates the two approaches; Sect. 6.3 examines the point empirically.

2.2 Positioning of the Proposed Framework

The preceding review identifies five requirements that an SDG evaluation framework must satisfy to address the problem set out in Sect. 1: heterogeneous multi-granular input (R1), non-additive aggregation (R2), structured capacity elicitation (R3), exact linguistic output (R4), and an operational tool (R5), each defined in the caption of Table 1. That table evaluates the principal families of existing approaches against them. The pattern of gaps is systematic rather than incidental: the approaches that handle representational heterogeneity aggregate additively, while those that model interaction assume a common cardinal domain and a single decision maker. No family satisfies R3 and R5 together with R1.

Two entries warrant comment. The MCDM-based SDG row is marked ∘ on R1 because fuzzy extensions admit imprecision within a single format but not several formats across criteria, and ∘ on R3 because protocols such as AHP elicit weights structurally, though for an additive model requiring no capacity. The DCM-based row satisfies R2 and R3 fully—it is the state of the art in capacity construction, and we adopt its elicitation front-end for that reason (Sect. 5.7)—but does not address the representational requirements R1 and R4, nor R5.

3 The Heterogeneous Fuzzy Linguistic Evaluation Model

This section formalises the computational framework underpinning the proposed SDG assessment model. The framework operationalises the Computing with Words (CWW) paradigm [34], enabling reasoning directly over linguistic concepts rather than

numerical abstractions. Operationally, CWW involves a translation phase—converting heterogeneous inputs into a unified internal representation—and a retranslation phase—converting aggregated results back into human-readable linguistic outputs [35–37].

The framework is structured around three aggregation levels operating over a unified 2-tuple representation:

- **Expert level:** For each criterion, the K individual assessments from the expert panel are consolidated into a single panel assessment using the 2-tuple weighted average or an OWA operator, depending on whether expert-specific weights are available or a positional aggregation is preferred.
- **Criterion level:** The per-criterion panel assessments are organised, filtered, and made available for inspection and partial aggregation—enabling, for instance, subgroup evaluations over subsets of criteria using the 2-tuple weighted average with domain-elicited importance weights.
- **Global level:** The full vector of criterion assessments is aggregated into a single overall SDG evaluation via the Choquet integral governed by a 2-additive capacity, capturing non-additive interactions among criteria.

The pipeline proceeds through four sequential pre-processing stages before reaching these aggregation levels: (i) standardisation of heterogeneous inputs, (ii) transformation into fuzzy sets over a common linguistic domain, (iii) conversion into linguistic 2-tuples within a shared Basic Linguistic Term Set, and (iv) multi-level aggregation.

3.1 Standardisation

To guarantee mathematical coherence across heterogeneous inputs representing cost and benefit criteria, all input data is normalised into the unit interval [0, 1], mapping the original range of values to a uniform scale. This prevents scale-induced biases during aggregation and ensures that all criteria are evaluated on an equal footing prior to semantic transformation [38].

3.2 Transformation into Fuzzy Sets

Normalised numerical information is unified into a high-granularity linguistic domain referred to as the *Basic Linguistic Term Set* (BLTS), denoted $\overline{S} = \{\overline{s}_0, \overline{s}_1, \ldots, \overline{s}_g\}$, strategically selected to preserve the maximum amount of initial knowledge [10]. The BLTS terms are assumed equidistant on [0, 1]: each term $\overline{s}_i$ is characterised by a triangular membership function

$$\mu_{\overline{s}_i}(u) = \max\left(0,\ 1 - g\ \left|u - \tfrac{i}{g}\right|\right), \quad u \in [0,1],\ i = 0, \ldots, g, \tag{1}$$

centred at i/g with support $[(i-1)/g,\ (i+1)/g] \cap [0, 1]$. This equidistant partition is standard in the 2-tuple literature [2, 10] and is justified by two considerations: (i) it guarantees a covering of [0, 1] in which every point belongs to at most two adjacent

membership functions, preserving a natural interpolation semantics, and (ii) the 2-tuple operators Δ and χ are defined on $[0, g]$ and rely on uniform spacing between successive linguistic labels for their rounding and centroid operations to be consistent. Non-uniform partitions would require modified Δ and χ functions and are outside the scope of the present framework.

Each normalised value is transformed into a fuzzy set on $\overline{S}$, denoted $\mathcal{F}(\overline{S})$, via the numerical transformation function $T_{N\overline{S}} : [0, 1] \to \mathcal{F}(\overline{S})$:

$$T_{N\overline{S}}(u) = \sum_{i=0}^{g} (\overline{s}_i / \gamma_i)$$

where $\gamma_i = \mu_{\overline{s}_i}(u) \in [0, 1]$ is the membership degree of u to the linguistic term $\overline{s}_i \in \overline{S}$.

Interval-valued assessments are unified analogously. Let $I = [\underline{u}, \overline{u}] \subseteq [0, 1]$ be a normalised interval estimate, characterised by its membership function $\mu_I(y) = 1$ if $y \in I$ and 0 otherwise. The interval transformation function $T_{I\overline{S}} : \mathcal{I}([0, 1]) \to \mathcal{F}(\overline{S})$ [10] is defined by the following:

$$T_{I\overline{S}}(I) = \sum_{i=0}^{g} (\overline{s}_i / \gamma_i), \quad \gamma_i = \max_{y \in [0,1]} \min\{\mu_I(y),\ \mu_{\overline{s}_i}(y)\}$$

so that each BLTS term receives the maximal compatibility between the interval and its membership function. Since $T_{I\overline{S}}$ reduces to $T_{N\overline{S}}$ when $\underline{u} = \overline{u}$, both numerical and interval inputs are handled by a single unification mechanism. The case study of Sect. 5 employs the numeric and linguistic domains only; interval-valued elicitation is supported by the framework and the tool without further modification.

This mechanism accommodates cross-format heterogeneity—numerical, interval-valued, and multi-granular linguistic assessments, with experts employing 3-term, 5-term, or 7-term scales—within a single computational pipeline [39, 40]. Analogous representational benefits have been demonstrated in complex performance evaluation [16] and supply-chain risk assessment [41].

3.3 Transformation into Linguistic 2-Tuples

To facilitate aggregation and produce continuous, interpretable results, the fuzzy sets obtained in the previous step are converted into linguistic 2-tuples [2]. The 2-tuple representation model encodes linguistic information via a pair (s, α), where $s \in S = \{s_0, \ldots, s_g\}$ is a base linguistic term and $\alpha \in [-0.5, 0.5)$ is a real-valued symbolic translation [2].

Definition 1 Let $S = \{s_0, \ldots, s_g\}$ be a set of linguistic terms. The *2-tuple set associated with S* is defined as $\langle S \rangle = S \times [-0.5, 0.5)$. The function $\Delta_S : [0, g] \longrightarrow \langle S \rangle$ [2] is defined by the following:

$$\Delta_S(\beta) = (s_i, \alpha), \quad \text{with} \quad \begin{cases} i = \text{round}(\beta), \\ \alpha = \beta - i, \end{cases} \tag{2}$$

where *round* assigns to β the closest integer $i \in \{0, 1, \ldots, g\}$.

The function Δ_S is bijective [2], with inverse $\Delta_S^{-1} : \langle S \rangle \to [0, g]$ defined by $\Delta_S^{-1}(s_i, \alpha) = i + \alpha$. This isomorphism identifies 2-tuples in $\langle S \rangle$ with values in the continuous interval $[0, g]$, avoiding the rounding loss incurred by purely symbolic operations.

Labelling convention. Throughout this paper, the linguistic label reported for a 2-tuple (s_i, α) is that of its base term s_i, as prescribed by the retranslation step of the 2-tuple model [2]; the symbolic translation α expresses the displacement of the assessment within (and towards the boundaries of) that term. Label transitions therefore occur exactly at half-integer values of $\beta = \Delta_S^{-1}(s_i, \alpha)$.

Fuzzy sets in $\mathcal{F}(\overline{S})$ are converted into linguistic 2-tuples via the function χ:

Definition 2 [10] Given $S = \{s_0, s_1, \ldots, s_g\}$, the function $\chi : \mathcal{F}(S) \longrightarrow \langle S \rangle$ is defined by the following:

$$\chi\left(\{\gamma_0, \gamma_1, \ldots, \gamma_g\}\right) = \Delta_S\left(\frac{\sum_{j=0}^{g} j\,\gamma_j}{\sum_{j=0}^{g} \gamma_j}\right) = (s_i, \alpha) \in \langle S \rangle. \tag{3}$$

Recent advances—including Hesitant Fuzzy Linguistic Term Sets (HFLTS) [25] and Probabilistic Linguistic Term Sets (PLTS) [33]—have further expanded the expressive capacity of this paradigm, allowing experts to express intra-personal hesitation between adjacent terms or distributional uncertainty over a term set. These representations offer richer individual expressiveness at the cost of substantially increased elicitation complexity, which conflicts with the requirements of large multidisciplinary panels: HFLTS requires experts to identify multiple acceptable terms per criterion, and PLTS requires probability assignments over full term sets—cognitive demands that are unrealistic for non-specialist evaluators such as geriatricians or health policymakers. The 2-tuple framework adopted here provides exact computation within the 2-tuple domain and a cognitively accessible single-label elicitation structure, making it the appropriate choice for heterogeneous panels where format accessibility and inter-format consistency are primary constraints. Extension to HFLTS-based elicitation for specialist subpanels constitutes a natural direction for future work.

3.4 Expert-Level Aggregation

The first aggregation level consolidates the K individual assessments provided for each criterion c_i into a single panel assessment. Since all assessments have been transformed into the common BLTS domain $\overline{S}$, the bijection $\Delta_{\overline{S}}^{-1}$ yields an exact numerical equivalent for each, enabling exact arithmetic in the continuous interval $[0, g]$ without further approximation.

2-Tuple Weighted Average (2t-WA) When expert-specific importance weights $\mathbf{w} = (w_1, \ldots, w_K)$, $\sum_k w_k = 1$, are available—for instance, from expert credibility scores or domain authority levels—the panel-aggregated 2-tuple for criterion c_i is as follows:

$$(s_i, \alpha_i) = \Delta_{\overline{S}}\left(\sum_{k=1}^{K} w_k \cdot \Delta_{\overline{S}}^{-1}(s_i^k, \alpha_i^k)\right) \tag{4}$$

When all experts are treated as equally reliable ($w_k = 1/K$), this reduces to the standard 2-tuple arithmetic mean [2], which is exact by the bijection property of $\Delta_{\overline{S}}$.

2-Tuple OWA Operator When a positional aggregation strategy is preferred—for instance, to apply an optimistic, neutral, or pessimistic stance—the 2-tuple Ordered Weighted Averaging (OWA) operator [2] is used instead. The assessments $\{\Delta_{\overline{S}}^{-1}(s_i^k, \alpha_i^k)\}$ are first sorted in the chosen order and then combined with a position-based weighting vector $\mathbf{v} = (v_1, \ldots, v_K)$:

$$(s_i, \alpha_i) = \Delta_{\overline{S}}\left(\sum_{k=1}^{K} v_k \cdot \beta_{(k)}\right) \tag{5}$$

where $\beta_{(k)}$ denotes the k-th value in the sorted sequence. Common choices include uniform weights ($v_k = 1/K$, equivalent to 2t-WA), linear pessimistic weights ($v_k \propto K + 1 - k$, emphasising low assessments), and linear optimistic weights ($v_k \propto k$).

The tool described in Sect. 4.2 exposes both operators and allows the analyst to select the appropriate strategy per panel and criterion.

3.5 Criterion-Level Results

At the criterion level, the per-criterion panel assessments (s_i, α_i) produced by the expert-level aggregation are made available for inspection, ranking, and partial aggregation over criterion subsets. If the analyst wishes to obtain a sub-dimensional evaluation—for instance, aggregating only infrastructure-related criteria—the 2-tuple weighted average (4) can be applied over the relevant subset with domain-elicited importance weights. This intermediate level supports diagnostic use: the criterion-level vector provides a fine-grained performance profile that is preserved alongside the global evaluation, enabling policy-relevant interpretation of which dimensions drive the overall result.

3.6 Global-Level Aggregation via Choquet Integral

The global level aggregates the criterion-level assessment vector $\mathbf{r} = \{(s_1, \alpha_1), \ldots, (s_n, \alpha_n)\}$ into a single overall SDG evaluation. We use the discrete Choquet integral [12, 42] governed by a 2-additive capacity (fuzzy measure), which generalises additive weighted averaging by capturing synergistic or redundant interactions among criteria—a key requirement in SDG evaluation, where inter-goal interdependencies are extensively documented [24].

Classical additive operators such as the 2-tuple Linguistic Weighted Average (2t-LWA) [2] assign a fixed scalar weight to each criterion independently, with no representation of interactions. The Choquet integral replaces this with a measure defined over all subsets of criteria, allowing the contribution of any group of criteria to differ from the sum of their individual contributions. Table 2 summarises the positioning of the proposed operator.

3.6.1 Choquet Integral and 2-Additive Capacities

Each criterion assessment (s_i, α_i) is converted to its normalised numerical equivalent:

$$\beta_i = \Delta_{\overline{S}}^{-1}(s_i, \alpha_i)/g \in [0, 1]$$

The discrete Choquet integral with respect to a fuzzy measure $\mu : 2^N \to [0, 1]$ (with $\mu(\emptyset) = 0$, $\mu(N) = 1$, and monotone) is as follows:

$$\mathcal{C}_\mu(\mathbf{r}) = \sum_{i=1}^{n} \beta_{(i)} \left[\mu(A_{(i)}) - \mu(A_{(i+1)})\right] \tag{6}$$

where $\beta_{(1)} \leq \cdots \leq \beta_{(n)}$ is the sorted sequence, $A_{(i)} = \{c_{(i)}, \ldots, c_{(n)}\}$, and $A_{(n+1)} = \emptyset$. The result is retranslated via the following:

$$\hat{r} = \Delta_{\overline{S}}\big(g \cdot \mathcal{C}_\mu(\mathbf{r})\big) \in \langle \overline{S} \rangle \tag{7}$$

To avoid specifying $\mu(A)$ for all 2^n subsets, we represent the capacity through its *Möbius transform* [11, 12] truncated at order 2—that is, a 2-*additive capacity*.

Table 2 Comparative analysis of aggregation operators relevant to heterogeneous SDG evaluation. Columns: 2t-LWA and 2t-IOWA [2, 4]; LC16 [31]; LC17 [32]; DCM-Ch. [13–15] (DCM-based Choquet capacity construction in MCDA); Proposed (this work). Rows compare: whether the operator accepts inputs in different representational formats (*Heterogeneous*); whether it models non-additive criterion dependencies (*Interaction*); whether it operates in the exact 2-tuple domain (*Exact 2-tuple*); whether it accepts multi-granular linguistic scales (*Multi-granular*); whether it satisfies formal monotonicity and idempotency; and whether it includes a structured elicitation protocol for the weight/density vector (*DCM*). The proposed framework is the only one combining DCM-based capacity elicitation with exact heterogeneous multi-granular 2-tuple unification

Property	2t-LWA	2t-IOWA	LC16	LC17	DCM-Ch.	Proposed
Heterogeneous input handling	✓	✓	✓	✓	×	✓
Criterion interaction modelling	×	×	×	✓	✓	✓
Exact 2-tuple computation	✓	✓	×	×	×	✓
Multi-granular linguistic input	✓	✓	×	×	×	✓
Formal monotonicity	✓	✓	✓	✓	✓	✓
Idempotency	✓	✓	✓	✓	✓	✓
Structured weight/density elicitation (DCM)	×	×	×	×	✓	✓

The Möbius representation $m : 2^N \to \mathbb{R}$ of a capacity μ is defined by $m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \mu(B)$, and recovers the capacity through $\mu(A) = \sum_{B \subseteq A} m(B)$. A capacity is *k-additive* when $m(B) = 0$ for every B with $|B| > k$; the 2-additive case retains only singleton and pairwise terms, so the analyst specifies just n *fuzzy densities* $g_i = m(\{i\})$ and a set of pairwise interaction masses $m(\{i, j\})$, all elicited from domain experts:

$$\mu(A) = \sum_{i \in A} g_i + \sum_{\{i,j\} \subseteq A} m(\{i, j\}), \qquad A \subseteq N \tag{8}$$

Equivalently, the tool accepts the pairwise terms as the *absolute* coalition values $\mu(\{i, j\})$, from which the Möbius mass follows as $m(\{i, j\}) = \mu(\{i, j\}) - g_i - g_j$. A positive $m(\{i, j\})$ encodes *synergy* between c_i and c_j (the pair contributes more than the sum of its densities), a negative one encodes *redundancy*, and $m(\{i, j\}) = 0$ for all pairs recovers the additive weighted average with weights g_i. The representation must satisfy the boundary and monotonicity conditions of a capacity, which for the 2-additive case reduce to the linear constraints of Lemma 1.

Lemma 1 (Well-posedness of the 2-additive capacity) *Let $g_i \geq 0$ and $m(\{i, j\}) \in \mathbb{R}$ be given. The set function defined by Eq. 8 is a normalised capacity (i.e. $\mu(\emptyset) = 0$, $\mu(N) = 1$, and $\mu(A) \leq \mu(B)$ whenever $A \subseteq B$) if and only if*

$$\sum_{i \in N} g_i + \sum_{\{i,j\} \subseteq N} m(\{i, j\}) = 1 \quad \textit{(normalisation)},$$

$$g_i + \sum_{j \in A} m(\{i, j\}) \geq 0 \quad \forall\, i \in N,\ \forall\, A \subseteq N \setminus \{i\} \quad \textit{(monotonicity)}.$$

In particular, if all masses are non-negative ($g_i \geq 0$, $m(\{i, j\}) \geq 0$) and normalised, both conditions hold automatically.

Proof Normalisation is Eq. 8 evaluated at $A = N$. For monotonicity, adding an element i to a coalition $A \not\ni i$ changes the capacity by $\mu(A \cup \{i\}) - \mu(A) = g_i + \sum_{j \in A} m(\{i, j\})$, which must be non-negative for all such A; taking the minimum over A yields the stated condition. When every mass is non-negative, the difference is a sum of non-negative terms, so monotonicity is immediate [11]. □

Relation to the λ-Measure and Choice of Parameterisation Relation to the lambda-measure and choice of parameterisation. A classical single-parameter alternative is the λ-fuzzy measure of Sugeno [43], in which a single scalar λ—the unique nonzero root of $1 + \lambda = \prod_i (1 + \lambda g_i)$—fixes *all* interactions from the densities alone, imposing a uniform synergy or redundancy across every coalition. This is more parsimonious (n parameters, no pairwise elicitation) but strictly less expressive: it cannot represent *selective* interactions, in which specific pairs of criteria interact while others remain independent. SDG 3 evaluation is precisely such a case—connectivity interacts strongly with the software-layer criteria that depend on it, but not with, say, data security. We therefore adopt the 2-additive representation, which localises interaction

to the pairs where domain knowledge supports it, at the cost of eliciting those pairwise terms explicitly (Sect. 5.7).

Why Pairwise Interactions Suffice Here The choice of $k = 2$ is a modelling decision and requires justification on three grounds. *First, elicitation burden.* A general capacity on n criteria requires $2^n - 2$ values; the 2-additive representation requires $n + \binom{n}{2}$, and the 3-additive one $n + \binom{n}{2} + \binom{n}{3}$. For the $n = 6$ criteria of Sect. 5 these are 62, 21 and 41 respectively. The additional 20 terms are third-order statements of the form "the joint effect of c_i, c_j and c_k exceeds what their pairwise relations imply," which experts find substantially harder to introspect than pairwise dependency questions and which no established protocol elicits at panel scale. *Second, the structure of the documented dependencies.* The interdependencies reported in the SDG literature [21, 24] and the dependencies identified in the target domain are dyadic in form—a service depends on an enabling condition—rather than irreducibly joint. A third-order term is warranted only when three criteria interact in a way not recoverable from their pairwise relations, and we have no domain evidence of such an effect here. *Third, empirical insensitivity.* Because the claim "pairwise suffices" is a modelling assumption rather than a theorem, we test it directly in Sect. 5.7.5: extending the capacity with a third-order mass on the connectivity cluster $\{c_1, c_4, c_5\}$ and sweeping it across its plausible range leaves both the criterion ranking and the global linguistic label unchanged.

The implication should nonetheless be stated plainly: a 2-additive capacity cannot represent joint effects irreducible to pairwise relations, and in a domain where such effects are documented the framework would understate them. Migration to $k \geq 3$ capacities is supported by the same Möbius machinery, and the tool already implements interactions up to order three.

3.6.2 Formal Properties

The aggregation of Sect. 3.6 is not the Choquet integral alone but the composition of three maps: the 2-tuple inverse that places the criterion assessments on a common cardinal scale, the integral itself, and the retranslation that returns the result to the linguistic domain. Since the reliability of the output depends on that composition rather than on any of its parts, we state its properties explicitly.

Definition 3 Let $\overline{S} = \{\overline{s}_0, \ldots, \overline{s}_g\}$ be the BLTS and μ a capacity on $N = \{1, \ldots, n\}$. The *2-tuple Choquet operator* $\widehat{\mathcal{C}}_\mu^{\overline{S}} : \langle \overline{S} \rangle^n \to \langle \overline{S} \rangle$ is

$$\widehat{\mathcal{C}}_\mu^{\overline{S}}(x_1, \ldots, x_n) = \Delta_{\overline{S}}\Big(g \cdot \mathcal{C}_\mu\Big(\tfrac{1}{g}\Delta_{\overline{S}}^{-1}(x_1), \ldots, \tfrac{1}{g}\Delta_{\overline{S}}^{-1}(x_n)\Big)\Big), \tag{9}$$

and $\langle \overline{S} \rangle$ is totally ordered by $x \preceq y \iff \Delta_{\overline{S}}^{-1}(x) \leq \Delta_{\overline{S}}^{-1}(y)$.

Proposition 1 *The operator $\widehat{\mathcal{C}}_\mu^{\overline{S}}$ of Definition 3 satisfies the following:*

(i) Closure. $\widehat{\mathcal{C}}_\mu^{\overline{S}}(x_1, \ldots, x_n) \in \langle \overline{S} \rangle$ *for every input, so the output is a valid 2-tuple of the same scale as the inputs.*

(ii) Exactness. $\Delta_{\overline{S}}^{-1}\left(\widehat{\mathcal{C}}_{\mu}^{\overline{S}}(x_1,\ldots,x_n)\right) = g\,\mathcal{C}_{\mu}(\boldsymbol{\beta})$ *exactly: the composition introduces no rounding at any stage, and the symbolic translation carries the full residual.*

(iii) Idempotence. *If* $x_1 = \cdots = x_n = x$ *then* $\widehat{\mathcal{C}}_{\mu}^{\overline{S}}(x,\ldots,x) = x$*, with exact recovery of both the base term and the translation.*

(iv) Boundedness. $\min_i x_i \preceq \widehat{\mathcal{C}}_{\mu}^{\overline{S}}(x_1,\ldots,x_n) \preceq \max_i x_i$.

(v) Monotonicity. *If* $x_i \preceq y_i$ *for all* i*, then* $\widehat{\mathcal{C}}_{\mu}^{\overline{S}}(\mathbf{x}) \preceq \widehat{\mathcal{C}}_{\mu}^{\overline{S}}(\mathbf{y})$.

(vi) Granularity consistency. *Let* $\overline{S}'$ *have granularity* $g' \neq g$ *and let* $\rho_{g\to g'}(x) = \Delta_{\overline{S}'}\left(\frac{g'}{g}\Delta_{\overline{S}}^{-1}(x)\right)$ *be the rescaling of a 2-tuple onto* $\overline{S}'$*. Then,*

$$\tfrac{1}{g'}\Delta_{\overline{S}'}^{-1}\left(\widehat{\mathcal{C}}_{\mu}^{\overline{S}'}(\rho_{g\to g'}(x_1),\ldots,\rho_{g\to g'}(x_n))\right) = \tfrac{1}{g}\Delta_{\overline{S}}^{-1}\left(\widehat{\mathcal{C}}_{\mu}^{\overline{S}}(x_1,\ldots,x_n)\right). \quad (10)$$

The normalised global evaluation is therefore invariant to the granularity of the BLTS, and so is the induced ordering of any set of alternatives or criteria evaluated on it.

The proof is given in Online Resource 1, Section S4, together with a fuller discussion of the two properties that carry most of the weight here.

The first is exactness. Property (ii) is what separates the 2-tuple pipeline from a purely ordinal one, in which every aggregation step must return a term of the scale, so that each step rounds and the losses accumulate across the architecture. The second is granularity consistency, and its scope should be stated precisely, since it is narrower than it first appears: what property (vi) leaves invariant is the *normalised* value, and hence the *ranking*, not the reported label—a scale with more terms partitions $[0, 1]$ more finely, so the same value may fall in a differently named term. Granularity therefore governs the resolution at which a result is communicated, not the substance of the aggregation that produced it; Sect. 5.8.5 verifies both halves of this behaviour numerically on the case study.

Together, these six properties license the three-level architecture set out below: because the composite operator is closed, exact and monotone on $\langle\overline{S}\rangle$, its output at one level is a valid input at the next, and the pipeline can be composed without the representational losses that motivated the model in the first place.

3.6.3 Complete Three-Level Pipeline

1. *Expert-level aggregation:* For each criterion c_i, consolidate the K panel assessments into (s_i, α_i) via 2t-WA (4) or OWA (5).
2. *Criterion-level inspection:* Examine, rank, and optionally sub-aggregate the per-criterion results using 2t-WA over criterion subsets of interest.
3. *Density and interaction elicitation, and measure construction:* Elicit the densities g_i from domain experts via the Deck-of-Cards procedure and the pairwise interaction terms $m(\{i, j\})$ from targeted expert judgement on criterion dependencies (Sect. 5.7); assemble the 2-additive capacity μ via (8).
4. *Global Choquet aggregation:* Apply (6)–(7) to obtain $\hat{r} \in \langle\overline{S}\rangle$.

4 A Comprehensive Framework and Decision Support Tool for SDG Evaluation

4.1 Methodological Overview

The proposed framework instantiates the heterogeneous fuzzy linguistic model of Sect. 3 as an operational SDG evaluation architecture. It accepts multi-format expert input—crisp numerical values, interval estimates, and multi-granular linguistic assessments—and produces interpretable linguistic evaluations at three granularity levels (expert, criterion, and global), without imposing representational homogeneity on the expert panel and with no additional numerical loss within the unified 2-tuple representation.

The framework is structured around five sequential steps, illustrated in Fig. 1:

1. **Data collection and normalisation:** Quantitative indicators—including statistical measures from official SDG reporting mechanisms [5]—are normalised to [0, 1] to eliminate scale bias across cost and benefit criteria. Qualitative indicators are collected using expert-selected linguistic scales of varying granularity (3-, 5-, or 7-term sets), preserving each evaluator's natural representational format.
2. **Transformation into fuzzy linguistic information:** Normalised numerical values and linguistic assessments are mapped onto the Basic Linguistic Term Set (BLTS) $\overline{S}$ via $T_{N\overline{S}}$ [10], converting all heterogeneous inputs into a common fuzzy representation $\mathcal{F}(\overline{S})$.
3. **Conversion to linguistic 2-tuples:** Fuzzy sets are converted into 2-tuples $(s_i, \alpha_i) \in \langle \overline{S} \rangle$ via χ [2], enabling exact symbolic arithmetic over the continuous interval $[0, g]$.
4. **Expert-level and criterion-level aggregation:** For each criterion c_i, the K individual 2-tuple assessments are consolidated via the 2-tuple weighted average (4) or OWA (5). The resulting per-criterion assessments are reported at the criterion level and made available for partial aggregation over criterion subsets.
5. **Global Choquet aggregation:** The criterion-level assessment vector is aggregated into a single overall SDG evaluation via the Choquet integral $\mathcal{C}_\mu$, with a prespecified 2-additive capacity—fuzzy densities $\mathbf{g}$ and pairwise interaction terms—elicited from domain experts. This stage captures non-additive interactions among SDG criteria.

4.2 Decision Support Tool

To make the framework directly accessible to practitioners, we developed a decision support tool that implements the complete pipeline end-to-end. The tool is designed for non-specialist users: analysts, policymakers, or evaluation coordinators who need to conduct SDG assessments without detailed knowledge of the underlying 2-tuple formalism.

Architecture The tool is structured around four functional modules corresponding to the pipeline stages, together with a capacity-construction module that parameterises

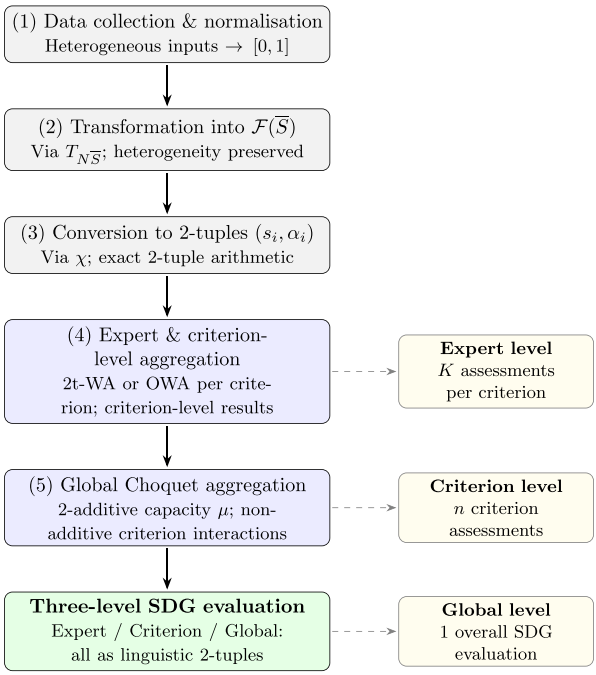


Fig. 1 Operational framework for SDG evaluation using the Heterogeneous Fuzzy Linguistic Model. Stages (4) and (5) (shaded) constitute the aggregation contribution. The framework produces results at three levels—expert, criterion, and global—all represented as interpretable linguistic 2-tuples

the global aggregation. Figure 2 shows the resulting system architecture and the data flow between modules.

1. **Panel configuration:** The analyst specifies the expert panel (K, professional groups, elicitation formats), the criteria set (n SDG indicators), and the BLTS granularity g. Format options include numeric scores on [0, 10], linguistic term sets of cardinality 3, 5, or 7, or mixed per-expert assignments. The tool validates format compatibility and enforces the granularity constraint $g \geq \max(|S^k| - 1)$.
2. **Assessment entry and unification:** Expert assessments are entered in their native format and automatically normalised and converted to 2-tuples in the BLTS domain. The unified assessment matrix is displayed alongside the original inputs, allowing the analyst to verify the transformation.
3. **Aggregation configuration:** For the expert level, the analyst selects the aggregation operator (2t-WA with optional expert weights, or OWA with a configurable weighting vector) per criterion. For the global level, the analyst enters the fuzzy densities **g** and the pairwise interaction terms, and the tool assembles the 2-additive capacity μ and executes the Choquet integral. Before proceeding, the tool verifies the normalisation and monotonicity conditions of Lemma 1; if any coalition violates monotonicity, the configuration is rejected and the offending coalition is reported, allowing the analyst to adjust the pairwise masses before continuing.

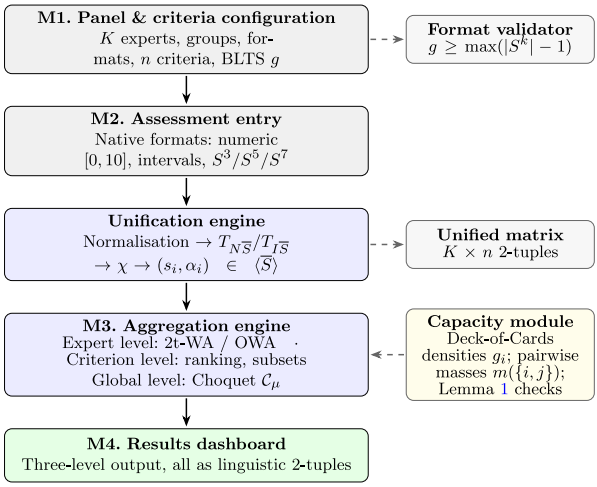


Fig. 2 System architecture and data flow of the decision support tool. Solid arrows denote the main processing pipeline; dashed arrows denote auxiliary data exchange. The four user-facing modules (M1–M4) are complemented by the capacity module, which supplies the 2-additive fuzzy measure that parameterises the global Choquet aggregation and verifies its well-posedness (Lemma 1) before the measure is applied

4. **Results dashboard:** The tool presents the three-level output: per-expert assessments per criterion (as the unified 2-tuple matrix), per-criterion panel assessments ranked by performance, and the global SDG evaluation. All results are displayed as linguistic 2-tuples with their natural-language labels (e.g. *Medium*, $(s_4, -0.10)$), and the criterion ranking is highlighted to identify structural bottlenecks.

Accessibility and Reproducibility The tool is implemented as a web application with a Java back-end (JDK 17, Apache Tomcat) serving JSP pages with an HTML front-end, deployable on any standard servlet container without additional dependencies. All computation is performed in closed form using the formulas of Sect. 3; the 2-additive capacity is assembled directly from its Möbius masses and no numerical optimisation is required.

4.3 Elicitation Protocol for the Interaction Structure

The fuzzy densities g_i come from the panel through the Deck-of-Cards procedure [17, 18]. The pairwise Möbius masses $m(\{i, j\})$ do not: they are the one input of the capacity the SRF procedure does not produce, so the framework specifies a protocol for them, set out in full in Online Resource 1, Section S8. In outline: interaction terms are elicited in three ordered stages—identification, with independence as the default so that a pair enters the structure only if a dependency is argued; sign, distinguishing synergy from redundancy; and magnitude, graded on a verbal scale anchored in advance rather than stated as a Möbius mass, which has no domain interpretation for the expert. An elicited structure is admissible only if it yields a capacity, which for the 2-additive case reduces to the checkable condition $g_i + \sum_{j \in A} m(\{i, j\}) \geq 0$ of Lemma 1, verified by the tool

before every aggregation; where densities are ranking-elicited, a second and tighter bound follows from requiring that the Shapley importances not invert the order the panel itself stated (Sect. 5.7). Disagreement is not resolved by forcing consensus: the median of the individual gradings is the point specification, and a pair whose gradings span more than one anchor is reported across its full elicited interval rather than at a point, so that a disputed interaction becomes an input whose consequences the analysis quantifies.

4.4 Framework Properties

The proposed framework offers three operationally significant advantages over existing SDG assessment architectures.

Representational Fidelity Prevailing SDG evaluation tools—including the SDG Index [8], the UN SDG Progress Dashboard [5], and composite indicator frameworks such as those reviewed in [19, 20]—operate exclusively on crisp numerical indicators and require all inputs to be pre-normalised onto a common scale before aggregation. This forces qualitative dimensions (institutional capacity, social equity, perceived quality of care) into artificial numerical proxies, introducing what [22] terms *measurement illusion*. MCDM-based extensions using fuzzy AHP or TOPSIS [9] partially address this by admitting fuzzy inputs, but still impose a fixed linguistic granularity on all experts, discarding the representational diversity of multidisciplinary panels. The proposed framework, by contrast, accepts heterogeneous formats natively and defers homogenisation to the mathematically exact 2-tuple transformation [2, 10], preserving each expert's judgement in its original granularity until the point at which exact 2-tuple conversion is formally guaranteed.

Multi-Level Diagnostic Output Existing SDG monitoring platforms—the UN Global SDG Indicators database [5], the SDG Tracker (Our World in Data), and the Sachs SDG Index [8, 21]—produce a single composite score or a flat indicator dashboard, with no structured aggregation hierarchy that relates expert judgments to criterion assessments to a global evaluation. Academic MCDM proposals for SDG evaluation [19, 20] similarly report final rankings or scores without exposing the intermediate diagnostic layers. The three-level architecture of the proposed framework—expert, criterion, and global—provides a complete and graduated picture of SDG performance: the expert level exposes inter-group divergence; the criterion level identifies structural bottlenecks; and the global level summarises overall progress in a single interpretable 2-tuple. This graduated view supports evidence-based policy analysis in a way that single-score evaluations cannot, and is directly operationalised by the tool described in Sect. 4.2.

Non-additive Criterion Interaction All major operational SDG frameworks—the UN composite indicators, the SDG Index, and the MCDM-based variants using weighted averages or AHP [9]—aggregate criteria additively, implicitly assuming that each SDG dimension contributes independently to overall sustainability performance. This

assumption conflicts with an extensive body of evidence on SDG interdependencies: [24] documents systematic synergies and trade-offs across all 17 goals, and [21] shows that progress on health goals is structurally contingent on infrastructure availability. The 2-additive capacity encodes synergistic or redundant interactions among criteria through expert-elicited pairwise interaction terms, capturing these systemic interdependencies in a parsimonious and formally grounded way that existing operational SDG evaluation tools do not currently offer.

5 Computational Verification, Simulation Experiment, and Illustrative Application to SDG 3

This section presents three complementary forms of empirical support for the proposed framework and decision support tool. Section 5.1 provides a *computational verification*: the code is executed under degenerate configurations whose analytical solution is known, confirming that the implementation correctly realises the equations of Sect. 3. Section 5.2 presents a *simulation experiment*: a controlled 3×2 factorial design that tests whether the Choquet aggregation produces the predicted discriminative patterns relative to additive baselines when the interaction structure and the position of the worst-performing criterion are varied by design. The remainder of the section presents a fully synthetic *illustrative application*—a simulated 50-expert multidisciplinary panel evaluating IoT-enabled healthcare technologies for ageing populations under SDG 3 (Good Health and Well-being)—designed to demonstrate the discriminative capacity of the Choquet aggregation relative to additive baselines under controlled conditions.

All three forms of support rest on synthetic data: no real experts participated at any stage, and the panel is generated by a seeded probabilistic protocol (Sect. 6.2.1). This is a deliberate design decision rather than a convenience. The object under study here is the *aggregation machinery*—whether the heterogeneous unification preserves the assessments it receives, and whether the 2-additive Choquet operator recovers a structural deficit that additive operators absorb. Establishing this requires knowing the ground truth of the data-generating process, which real panel data cannot supply: with a real panel the behaviour of the operator would be confounded with unobservable properties of the panel itself—its composition, its consensus dynamics, the unknown true profile it is assessing—and no discrepancy between operators could be attributed unambiguously to the aggregation step. A seeded generator fixes the criterion profile, the inter-group divergence structure and the granularity assignment by construction, so that every displacement in Table 11 is traceable to the operator under test, and makes the whole chain exactly reproducible—something a real elicitation panel cannot offer.

The counterpart is equally deliberate: the case study establishes *internal* validity only, and external validity is deferred to future work (Sects. 6 and 7). Table 3 states that boundary claim by claim. Every result reported in this section falls in the left-hand column; no statement in the right-hand column is asserted anywhere in the paper.

The choice of IoT-enabled healthcare for ageing populations as an illustrative domain is motivated by several recent developments. Digital health solutions for age-

ing populations are expanding rapidly, with telemedicine, wearable devices, and smart home technologies among the most active research areas [44–46]. However, rural communities continue to face systemic barriers—limited connectivity, low digital literacy, and inadequate infrastructure—that constrain telemedicine adoption [47–49]. Digital equity in healthcare remains an open challenge: the WHO Global Strategy on Digital Health 2020–2025 explicitly calls for frameworks that reduce rather than exacerbate healthcare disparities through digitisation [50–53]. Finally, the security and privacy challenges inherent to IoT-enabled medical devices add a further evaluative dimension that interacts non-additively with infrastructure and clinical criteria [54–57].

This application domain is particularly suited to the framework: it involves a genuinely multidisciplinary panel, combines quantitative and qualitative indicators that resist homogeneous representation, and concerns a sustainability dimension where infrastructure deficits and software capabilities interact non-additively. The entire evaluation was conducted using the tool described in Sect. 4.2, which produced the three-level output—expert, criterion, and global—reported here.

The synthetic scenario models a generic territorial context characterised by a dispersed ageing population and limited rural connectivity—conditions under which decentralised ICT infrastructure becomes a critical and contested sustainability variable, motivating both the scenario design and the Choquet-based aggregation framework proposed in this work. All expert assessments and elicitation parameters are synthetically generated under the protocol of Sect. 6.2.1 and reported in full in Online Resource 1 (Tables S2–S3); the purpose of the illustrative application is to test the mathematical robustness, heterogeneity handling, and discriminative capacity of the framework and the tool under controlled and fully reproducible conditions, not to report empirical findings about any real deployment.

Table 3 Scope of the synthetic case study. The distinction is between the behaviour of the aggregation machinery, observable under a known data-generating process, and the state of the world, which is not

What the case study establishes	What it does not establish
That the implementation correctly realises the equations of Sect. 3, including the degenerate identities verified in Sect. 5.1	That any real territory, population, or deployment attains the reported evaluation, or any other
That the heterogeneous unification carries numeric, interval-valued and multi-granular linguistic assessments into a common domain without the rounding loss that ordinal models incur	That the criterion set, the professional groups, or the granularity assignments used here reproduce those of any real SDG 3 evaluation panel
That the 2-additive Choquet operator recovers a structural deficit that additive operators absorb, and attributes the resulting displacement to a named criterion	That Rural Connectivity is a bottleneck for SDG 3 delivery in any real setting; the deficit is planted in the generator (Sect. 6.2.1)
That the aggregation conclusion is stable across the elicited interaction strength and across the scenario variations of Sect. 5.2	That the elicited densities or interaction terms reflect the preferences of any real panel; both were supplied by the research team
That every reported figure is exactly reproducible from the released seeded generator and analysis code	That the tool is usable by practitioners, which requires the evaluation deferred in Sect. 7

5.1 Computational Verification

Before applying the framework to the illustrative scenario, we verify its implementation against five degenerate configurations whose analytical solution is known.

1. **Additive capacity:** Setting all pairwise interaction terms to zero ($m(\{i, j\}) = 0$ for all pairs) reduces the 2-additive Choquet integral to a weighted average with weights equal to the normalised fuzzy densities $g_i / \sum_j g_j$. The tool output matches the closed-form weighted average to machine precision.
2. **Homogeneous panel:** When all K experts use the same linguistic scale (i.e. scale cardinality equal to that of the BLTS), the multi-granular unification reduces to the identity transformation. The tool output confirms that $\chi \circ T_{N\bar{S}}$ acts as the identity on inputs already expressed in the BLTS.
3. **Unanimous assessments:** When all experts assign the same label s_j for a given criterion, the expert-level 2-tuple weighted average returns $(s_j, 0)$ regardless of the weighting vector. The tool output matches for all criteria.
4. **Known capacity recovery:** A 2-additive capacity with three criteria ($g = (0.3, 0.3, 0.2)$, $m(\{1, 2\}) = m(\{2, 3\}) = 0.1$) and criterion values $(0.2, 0.5, 0.8)$ yields $\mathcal{C}_\mu = 0.2 \times 0.4 + 0.5 \times 0.4 + 0.8 \times 0.2 = 0.44$ by hand. The tool returns 0.44 exactly.
5. **Additive identity across random panels:** Check 1 verifies the identity on a single configuration; here it is stressed over 100 randomly generated heterogeneous panels with randomly drawn densities. In every replication the Choquet integral and the density-weighted average coincide ($\bar{\delta} = 0.000$, sd $= 0.000$), confirming that the implementation introduces no spurious interaction effect for any input configuration.

All five checks are passed exactly (relative error $< 10^{-12}$). The verification scripts and expected outputs are included in the replication package.

5.2 Simulation Experiment

Beyond verifying that the code implements the equations correctly, we conduct a controlled simulation experiment to test whether the framework behaves as theoretically expected under systematically varied conditions. This addresses a distinct question: not whether the code is correct, but whether the Choquet aggregation produces the predicted discriminative patterns relative to additive baselines when the interaction structure and the position of the worst-performing criterion are manipulated by design.

5.2.1 Experimental Design

We use a 3×2 factorial design with two factors:

- **Factor A—interaction structure:** (A1) no interaction (additive capacity, $m(\{i, j\}) = 0$ for all pairs); (A2) positive synergy ($m(\{i, j\}) = +0.05$ for one criterion pair); (A3) negative redundancy ($m(\{i, j\}) = -0.05$ for one pair).

- **Factor B—position of worst criterion:** (B1) the worst-performing criterion participates in the interaction pair; (B2) the worst-performing criterion does not participate in any interaction; the interacting pair is formed by the two criteria with the next-lowest indices after the worst, ensuring a deterministic and reproducible assignment.

Under condition A1, factors A and B are orthogonal (no interaction exists), so A1×B1 = A1×B2 by construction, yielding five distinct cells. Each cell is replicated $R = 100$ times. In every replication, a fresh synthetic panel of $K = 50$ experts with high heterogeneity (three groups using S^3, S^5, and S^7) is generated: true criterion values are drawn from Uniform(1, 7) on the BLTS, and expert assessments are sampled around these values with Gaussian noise ($\sigma = 1.5$), quantised to each expert's native scale, and unified to the BLTS. Densities are drawn from a randomised Deck-of-Cards procedure and normalised so that $\sum g_i + \sum m(\{i, j\}) = 1$.

Two auxiliary (non-factorial) conditions are also tested under the A2×B1 configuration:

- **Panel heterogeneity:** high (three scales) vs. low (all experts on S^7).
- **Panel size:** $K \in \{10, 30, 50\}$.

The response variable is the *Choquet–additive gap*: $\delta = \hat{\beta}^{\text{Choquet}} - \hat{\beta}^{\text{additive}}$, where the additive baseline uses the same singleton densities with all $m(\{i, j\}) = 0$.

5.2.2 Hypotheses

The Choquet–additive identity under an additive capacity (condition A1) is verified in Sect. 5.1 (check 5, additive identity); here, it serves as the zero-gap control row. The simulation tests three behavioural predictions:

H1. Under A3 (negative redundancy), the gap is positive ($\delta > 0$): the redundant pair contributes less jointly than the sum of its densities, raising the Choquet score relative to the additive baseline.
H2. Under A2 (positive synergy), the gap is largest in absolute value when the worst criterion participates in the synergy pair (A2×B1), because the superadditive capacity assigns the bottleneck criterion increased effective weight $\Delta\mu_i > g_i$.
H3. The variance of the global result decreases with panel size K, reflecting the averaging effect of larger panels.

5.2.3 Results

Table 4 reports the factorial results. The A1 rows confirm the zero-gap identity; the three hypotheses are assessed below.

Control (A1) Under the additive capacity, the Choquet integral and the density-weighted average coincide to machine precision ($\bar{\delta} = 0.000$, sd $= 0.000$) in every replication, as expected from the algebraic identity verified in Sect. 5.1 (check 5).

H1 is confirmed: under negative redundancy (A3), the gap is positive ($\bar{\delta} = +0.103$ for B1, $+0.024$ for B2; 95 % CI for B1: $[+0.097, +0.109]$), meaning that the Choquet

integral assigns a *higher* global score than the additive baseline. This is the expected behaviour: when two criteria are redundant, their joint contribution under the capacity is less than the sum of their densities, redistributing effective weight toward the remaining criteria and raising the global score when those criteria perform above the pair average.

H2 is confirmed: the absolute gap under positive synergy is five times larger when the worst criterion participates in the pair ($|\bar{\delta}| = 0.111$ for A2×B1 vs. 0.021 for A2×B2; 95 % CIs [0.104, 0.117] and [0.012, 0.030], non-overlapping; Cohen's $d = 2.24$). This confirms that the Choquet bottleneck-detection mechanism is driven by the interaction between the superadditive capacity structure and the position of the worst-performing criterion in the sorting order—the diagnostic property that additive operators structurally cannot exhibit.

H3 is confirmed by the auxiliary panel-size analysis (Table 5). Under the random-truth protocol (same ground-truth generation as the factorial design), the variability of the gap decreases modestly ($\mathrm{sd}(\delta) = 0.034 \to 0.033 \to 0.030$), but the global-score variability does not decrease monotonically ($\mathrm{sd}(\hat{\beta}^{\mathrm{Ch}}) = 0.747 \to 0.668 \to 0.705$), because the between-replication variance of the randomly drawn ground truth dominates. Under the fixed-truth protocol, which isolates the averaging effect by holding the true criterion values constant, both quantities decrease monotonically ($\mathrm{sd}(\delta) = 0.025 \to 0.016 \to 0.015$; $\mathrm{sd}(\hat{\beta}^{\mathrm{Ch}}) = 0.281 \to 0.221 \to 0.188$). The mean gap itself is stable across panel sizes ($-0.122, -0.115, -0.116$), indicating that panel size affects the precision of the estimate but not its central tendency.

The auxiliary heterogeneity analysis shows that the gap magnitude is stable across heterogeneity levels ($\bar{\delta} = -0.107 \pm 0.034$ under high heterogeneity vs. -0.113 ± 0.032 under low heterogeneity for A2×B1), indicating that the Choquet displacement is driven by the capacity structure rather than by the representational diversity of the panel—a reassuring property for practical deployment, where heterogeneity is a design feature of the framework rather than a confound.

Table 4 Simulation experiment: Choquet–additive gap $\delta = \hat{\beta}^{\mathrm{Choquet}} - \hat{\beta}^{\mathrm{additive}}$ across factorial conditions ($R = 100$ replications, $K = 50$, high heterogeneity). Condition A1×B2 is identical to A1×B1 by construction. The auxiliary analyses of Table 5 and Sect. 5.2.4 draw independent random streams from the same seeded generator, so their A2×B1 estimates differ from the corresponding factorial cell within sampling error

Factor A	Factor B	$\bar{\delta}$	$\mathrm{sd}(\delta)$	95 % CI	$\bar{\hat{\beta}}^{\mathrm{Ch}}$	$\bar{\hat{\beta}}^{\mathrm{Add}}$
A1: Additive	B1 (in pair)	−0.000	0.000	[0, 0]	4.062	4.062
A1: Additive	B2 (not in pair)	−0.000	0.000	[0, 0]	4.062	4.062
A2: Synergy	B1 (in pair)	−0.111	0.034	[−0.117, −0.104]	3.908	4.019
A2: Synergy	B2 (not in pair)	−0.021	0.045	[−0.030, −0.012]	3.893	3.914
A3: Redundancy	B1 (in pair)	+0.103	0.031	[+0.097, +0.109]	4.015	3.912
A3: Redundancy	B2 (not in pair)	+0.024	0.049	[+0.014, +0.034]	4.015	3.991

Table 5 Effect of panel size on the Choquet–additive gap and global score variability (A2×B1, high heterogeneity, $R = 100$). Under *random truth* (same as the factorial), fresh true values are drawn in each replication; under *fixed truth* (seed 7), only panel sampling varies

Protocol	K	$\bar{\delta}$	sd(δ)	$\bar{\hat{\beta}}^{\mathrm{Ch}}$	sd($\hat{\beta}^{\mathrm{Ch}}$)
Random truth	10	−0.109	0.034	3.908	0.747
	30	−0.108	0.033	3.868	0.668
	50	−0.100	0.030	3.962	0.705
Fixed truth	10	−0.122	0.025	4.561	0.281
	30	−0.115	0.016	4.552	0.221
	50	−0.116	0.015	4.560	0.188

5.2.4 Additional Scenarios

To test the generality of the Choquet displacement mechanism beyond the configuration of the factorial design, three further scenarios are evaluated using the same simulation generator and protocol as Sect. 5.2.1 ($R = 100$, $K = 50$, high heterogeneity, $|m(\{i, j\})| = 0.05$). They differ only in how the singleton densities are assigned and which criterion pair carries the interaction.

1. **Base (reference):** The synergy is placed on a pair containing the worst-performing criterion, and densities are assigned at random, as in condition A2×B1.
2. **Low-density bottleneck:** The synergy pair is unchanged, but densities are assigned in inverse rank order, so the worst-performing criterion receives the *smallest* singleton density and the best-performing criterion the largest.
3. **Non-bottleneck interaction:** The synergy is placed on the pair formed by the two *best*-performing criteria, so that the worst criterion participates in no interaction; densities are assigned at random as in the base case.

Table 6 confirms three properties of the Choquet displacement. First, the base-case gap ($\bar{\delta} = -0.106$) reproduces condition A2×B1 of the factorial design, as expected. Second, the gap is *larger* in absolute value when the bottleneck criterion carries the lowest singleton density ($|\bar{\delta}| = 0.122$ vs. 0.106): the relative amplification $\Delta\mu_i / g_i$ grows as g_i shrinks, so the superadditive mechanism does not require the bottleneck to also be the most important criterion—it operates through the coalition structure rather than through the marginal densities. Third, when the synergy is moved to the pair of best-performing criteria (Scenario 3), the gap *reverses sign* ($\bar{\delta} = +0.061$, 95 % CI [+0.056, +0.067]): the Choquet integral now amplifies criteria that already perform above the panel average, raising the global score above the additive baseline. This sign

Table 6 Additional simulation scenarios ($R = 100$, $K = 50$, high heterogeneity). The gap δ is defined as in Table 4. The sign reversal in Scenario 3 confirms that the Choquet displacement is governed by the interaction between the capacity structure and the position of the interacting criteria in the sorted order

Scenario	$\bar{\delta}$	sd(δ)	95% CI	$\bar{\hat{\beta}}^{\mathrm{Ch}}$	$\bar{\hat{\beta}}^{\mathrm{Add}}$
1. Base (worst criterion in pair)	−0.106	0.031	[−0.112, −0.100]	3.942	4.048
2. Low-density bottleneck	−0.122	0.033	[−0.128, −0.115]	4.249	4.371
3. Non-bottleneck interaction	+0.061	0.029	[+0.056, +0.067]	4.109	4.048

reversal confirms that the direction of the displacement is determined by the interaction between the capacity structure and the rank position of the interacting criteria—the core diagnostic property that additive operators cannot reproduce.

The complete simulation code, seeded random generator, and raw output are included in the replication package.

5.2.5 Strengthened Validation

Three further analyses strengthen the internal validation, all at 1000 replications per cell with a fixed seed and reported in full in Online Resource 1, Section S6. Percentile bootstrap confidence intervals on the Choquet–additive gap exclude zero in all four interacting conditions of the factorial design, while the two non-interacting conditions reproduce the additive result exactly. Extending the panel-size sweep to $K \in \{10, 25, 50, 100, 200\}$ under a fixed ground truth shows the gap converging monotonically with the standard deviation halving, so panel size governs the precision of the estimate and not the effect. Finally, a recovery study asks whether the pipeline returns what the data-generating process put in—a question no real panel can answer, since its ground truth is unobservable, so the synthetic design is here not a weaker substitute for real data but the only setting in which the question is well posed. Planting a deficient criterion and running the pipeline blind recovers it in 100% of replications from $K = 25$ upward when the deficit is clearly separated, and degrades gracefully to 75.1% when the planted criterion sits only 0.25 BLTS units below its nearest neighbour, below the expert noise. The estimation error is constant across that range (MAE ≈ 0.186), which locates the failures in genuine overlap between criteria rather than in bias introduced by the aggregation, and bounds what the framework can be asked to do: deficits smaller than the dispersion of the panel are not reliably identifiable by any procedure operating on that panel.

5.3 Decision-making Framework

The evaluation is formalised as an MCDM problem [7, 58] over the following sets:

Simulated panel $E = \{e_1, \ldots, e_{50}\}$: a synthetic multidisciplinary panel of $|E| = 50$ evaluators comprising 15 geriatricians (e_1–e_{15}, group G), 20 IT engineers (e_{16}–e_{35}, group IT), and 15 health policymakers (e_{36}–e_{50}, group P), representing the professional profiles typically involved in IoT-enabled healthcare evaluation. Expert assessments were synthetically generated via the seeded structured simulation protocol of Sect. 6.2.1, with parameter distributions designed to reproduce the representational heterogeneity and inter-group divergence characteristic of real multidisciplinary panels. This fully synthetic design enables exact replication and controlled sensitivity testing of the framework under stress conditions that would be difficult to isolate in field data.

Criteria $C = \{c_1, \ldots, c_6\}$: six SDG 3 indicators for IoT-enabled health:

- c_1: Telemedicine adoption
- c_2: Device access and affordability
- c_3: Home accessibility and sensorization
- c_4: Rural connectivity

- c_5: Treatment adherence tracking
- c_6: Data security and privacy

Elicitation Formats Two information domains are considered. The **numeric domain** D_N uses integer scores on [0, 10]; the **linguistic domain** D_L uses multi-granular term sets S^3, S^5, S^7 as defined below:

$$S^3 = \{\text{Low},\ \text{Medium},\ \text{High}\}$$
$$S^5 = \{\text{Very Low},\ \text{Low},\ \text{Medium},\ \text{High},\ \text{Very High}\}$$
$$S^7 = \{\text{Very Low},\ \text{Low},\ \text{Med-Low},\ \text{Medium},\ \text{Med-High},\ \text{High},\ \text{Very High}\}$$

Format assignment follows the nature of each criterion and the domain knowledge of each group. The information domain is a property of the *criterion*: criteria with objectively measurable indicators are assessed numerically by all experts, whereas inherently judgmental criteria are assessed linguistically by all experts. The granularity of the linguistic scale, in turn, is a property of the *expert group*: groups with deeper domain knowledge of a criterion are assigned finer scales (S^7), while groups with limited knowledge use coarser scales (S^5 or S^3), on the rationale that experts who can discriminate more finely between adjacent levels should be given the representational means to do so. This assignment reflects representational capacity, not aggregation weight: all experts receive equal weight ($w_k = 1/K$) in the expert-level aggregation (Sect. 5.6). In applications where differential credibility is warranted, the framework supports non-uniform w_k (4), and assigning higher weight to experts using finer scales would be a coherent design choice—one that could be explored in future work with real panels. Concretely, c_4 (Rural Connectivity) and c_5 (Treatment Adherence Tracking) are numeric for all groups—both admit direct quantitative indicators (coverage and bandwidth availability; recorded adherence rates)—while the remaining four criteria are linguistic with group-specific granularities: c_1 (G: S^7, IT: S^5, P: S^5), c_2 (G: S^5, IT: S^7, P: S^5), c_3 (G: S^7, IT: S^3, P: S^5), and c_6 (G: S^3, IT: S^7, P: S^5). Thus, clinicians discriminate finely on care-related criteria (c_1, c_3) but coarsely on data security (c_6), whereas IT engineers do the opposite (c_2, c_6 vs. c_3). This design encodes the two forms of representational heterogeneity the framework must handle: numeric vs. linguistic information across criteria, and multi-granular linguistic information across experts.

Fuzzy densities $\mathbf{g} = (g_1, \ldots, g_6)$—encoding the importance of each criterion under SDG 3—were constructed through a Deck-of-Cards procedure [17] conducted as part of the synthetic study design (Sect. 5.7), and are reported in Table 8.

Basic Linguistic Term Set All inputs are unified onto the Basic Linguistic Term Set $\overline{S} = \{s_0, s_1, \ldots, s_8\}$ ($g + 1 = 9$ terms, $g = 8$):

$$s_0 : \text{Null} \quad s_1 : \text{Very Low} \quad s_2 : \text{Low} \quad s_3 : \text{Low-Med}$$
$$s_4 : \text{Medium} \quad s_5 : \text{Med-High} \quad s_6 : \text{High} \quad s_7 : \text{Very High}$$
$$s_8 : \text{Complete}$$

The granularity $g = 8$ satisfies $g \geq \max(|S^k| - 1)$ for all linguistic elicitation scales, ensuring that the BLTS has sufficient cardinality to host every source scale [10]. In terms of positional alignment, S^3 and S^5 nest exactly onto integer positions of the BLTS ($S^3 \mapsto \{0, 4, 8\}$; $S^5 \mapsto \{0, 2, 4, 6, 8\}$), because g is divisible by their respective denominators ($|S^k| - 1 \in \{2, 4\}$). S^7, however, maps onto fractional positions ($k/6 \times 8 \in \{0, 1.33, 2.67, 4, 5.33, 6.67, 8\}$), since $6 \nmid 8$. This does not introduce numerical error within the 2-tuple representation: the symbolic translation α encodes the fractional displacement exactly, and all subsequent aggregation operates on the continuous interval $[0, g]$ via $\Delta_{\overline{S}}^{-1}$. It does mean, however, that four of the seven S^7 terms do not coincide with base labels of the BLTS, so their retranslated label inherits a nonzero α even before aggregation. A BLTS with $g = 12$ ($\text{lcm}(2, 4, 6) = 12$, 13 terms) would place all three scales on integer positions, at the cost of a finer granularity than is needed for interpretive reporting; the present choice of $g = 8$ balances alignment coverage with label parsimony and is the standard recommendation in the 2-tuple literature [2, 10]. The extreme labels *Null* and *Complete* denote the absence and full achievement of the assessed criterion, respectively; they do not appear in any elicitation scale (S^3, S^5, S^7) and serve only as the boundary anchors of the BLTS required by the 2-tuple formalism. We prefer *Complete* over *Total* to avoid connotation with a numerical total or sum.

Figure 3 shows the panel and criteria configuration screens of the tool as used in this case study.

5.4 Heterogeneous Evaluation Matrix

The complete heterogeneous assessment matrix for all 50 experts is reported in Online Resource 1 (Table S2). Each cell reports either a numeric score $v \in \{0, \ldots, 10\}$ or a linguistic label with its scale (e.g. *Medium (L5)*), as provided natively by each evaluator.

5.5 Unification into the BLTS

5.5.1 Transformation Rules

Numeric inputs $v \in [0, 10]$ are mapped to $\beta \in [0, 8]$ via the following:

$$\beta = \frac{v}{10} \times g, \quad g = 8 \tag{11}$$

Linguistic inputs at index $k \in \{0, \ldots, n - 1\}$ in a scale of size n are mapped via the following:

$$\beta = \frac{k}{n - 1} \times g \tag{12}$$

Each β is then encoded as a 2-tuple via $\Delta_{\overline{S}}(\beta) = (s_i, \alpha)$ (Definition 1).

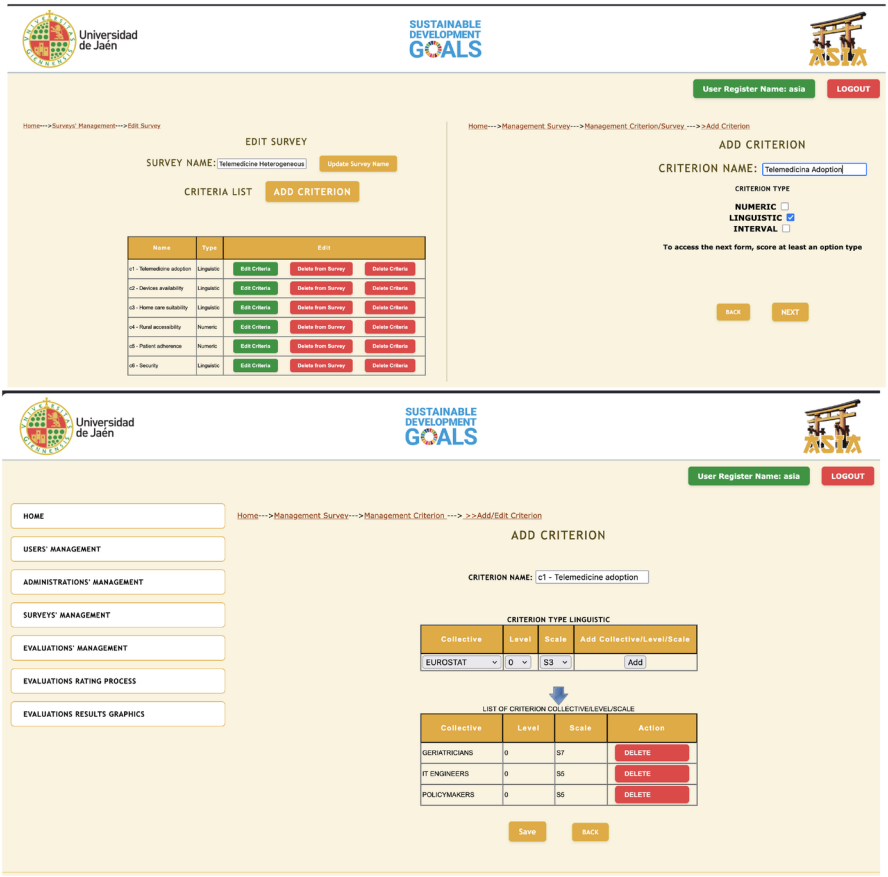


Fig. 3 Tool screen 1a: criteria list of the SDG 3 evaluation. The six criteria are declared as numeric (c_4, c_5) or linguistic (the remaining four) and Tool screen 1b: per-criterion elicitation formats by expert group. For criterion c_1 (Telemedicine Adoption), the scale assigned to each group is S^7 for geriatricians and S^5 for IT engineers and policymakers, as specified in Sect. 5.3

5.5.2 Illustrative Examples

Example 1 **Numeric input** Expert e_{16} (IT) evaluates c_5 (Adherence) with $v = 4$:

$$\beta = \frac{4}{10} \times 8 = 3.20 \;\Rightarrow\; \Delta_{\overline{S}}(3.20) = (s_3,\ +0.20)$$

Example 2 **Linguistic input,** S^5 Expert e_{16} (IT) evaluates c_1 (Telemedicine) as *High* in S^5 (index $k = 3$, $n = 5$):

$$\beta = \frac{3}{4} \times 8 = 6.00 \;\Rightarrow\; \Delta_{\overline{S}}(6.00) = (s_6,\ 0.00)$$

Example 3 Linguistic input, S^7 Expert e_1 (G) evaluates c_1 (Telemedicine) as *Med-Low* in S^7 (index $k = 2, n = 7$):

$$\beta = \frac{2}{6} \times 8 = 2.67 \;\Rightarrow\; \Delta_{\overline{S}}(2.67) = (s_3,\ -0.33)$$

Example 4 Linguistic input, S^3**.** Expert e_1 (G) evaluates c_6 (Security) as *Medium* in S^3 (index $k = 1, n = 3$):

$$\beta = \frac{1}{2} \times 8 = 4.00 \;\Rightarrow\; \Delta_{\overline{S}}(4.00) = (s_4,\ 0.00)$$

Example 5 Interval-valued input Although not used in the present case study, the framework supports interval-valued assessments via $T_{I\overline{S}}$ (Sect. 3.2). Consider a hypothetical normalised interval $I = [0.30, 0.50]$. Using the triangular membership functions of Eq. 1 with $g = 8$, the maximal compatibility $\gamma_i = \max_{y\in[0,1]} \min\{\mu_I(y), \mu_{\overline{s}_i}(y)\}$ yields non-zero memberships for $\overline{s}_2$ through $\overline{s}_4$—specifically $\gamma_2 = 0.60$, $\gamma_3 = 1.00$ and $\gamma_4 = 1.00$—with centroid:

$$\begin{aligned}\beta = \chi(\{\gamma_0, \ldots, \gamma_8\}) &= \frac{2(0.60) + 3(1.00) + 4(1.00)}{0.60 + 1.00 + 1.00} \\ &= 3.154 \;\Rightarrow\; \Delta_{\overline{S}}(3.154) = (s_3,\ +0.15).\end{aligned}$$

5.5.3 Complete Unified Dataset

The full BLTS 2-tuple representation for all 50 experts across all six criteria is reported in Online Resource 1 (Table S3), obtained by applying (11)–(12) and Definition 1.

5.6 Expert-Level Aggregation

For each criterion c_i, the panel-aggregated 2-tuple is computed via the 2t-WA with uniform weights (4, $w_k = 1/K$), as no differential expert credibility scores are available in this study. The tool also supports OWA aggregation at this stage; uniform weights are used here for reproducibility. This uniform-weight 2t-WA governs only the expert-level aggregation; the criterion-level and global aggregations use the 2t-WA across criteria and the 2-additive Choquet integral, respectively (Sect. 3.6.1).

Figure 4 shows the assessment entry and unification screen of the tool, displaying the heterogeneous input matrix alongside the automatically computed BLTS 2-tuple representation for each expert and criterion.

Illustrative Computation for c_4 (Rural Connectivity) From the unified dataset (Online Resource 1, Table S3), the 50 normalised values $\beta_4^k = \Delta_{\overline{S}}^{-1}(s_4^k, \alpha_4^k) \in [0, 8]$ have mean $\bar{\beta}_4 = 2.400$, that is $\hat{\beta}_4 = 2.400/8 = 0.3000$ once normalised, yielding as the panel-aggregated 2-tuple of c_4:

$$\left(s^{(4)}, \alpha^{(4)}\right) = \Delta_{\overline{S}}(2.400) = (s_2, +0.40)$$

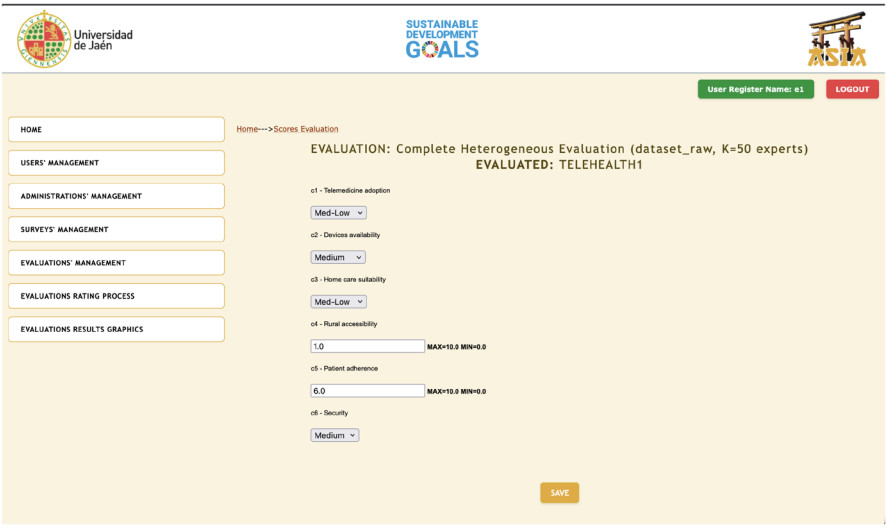


Fig. 4 Tool screen 2: assessment entry and automatic unification. Left panel shows expert assessments in native format (numeric scores and linguistic labels); the unified 2-tuple representation in the BLTS domain produced automatically by the tool

Table 7 summarises the expert-level aggregation results for all six criteria, constituting the criterion-level output reported by the tool.

5.7 2-Additive Capacity Construction

5.7.1 Deck-of-Cards Elicitation of the Fuzzy Densities

The framework prescribes the elicitation of the fuzzy densities $\mathbf{g}$, prior to evaluation, through the Deck-of-Cards Method in its Simos–Roy–Figueira (SRF) form [17], chosen for its low cognitive burden on non-specialist panels and its established track record in weight construction for MCDA models [18]. Recent extensions of the Deck-

Table 7 Panel aggregation results per criterion. $\bar{\beta}_i$ is the panel-aggregated assessment in BLTS units $[0, g]$ with $g = 8$; $\hat{\beta}_i = \bar{\beta}_i / g$ is the corresponding normalised value in $[0, 1]$ used as input to the Choquet integral (6)

Criterion	$\bar{\beta}_i$	$\hat{\beta}_i = \bar{\beta}_i / g$	Panel 2-tuple	Linguistic label
c_1 Telemedicine	4.440	0.5550	$(s_4, +0.44)$	Medium
c_2 Devices	4.107	0.5133	$(s_4, +0.11)$	Medium
c_3 Home	3.680	0.4600	$(s_4, -0.32)$	Medium
c_4 Rural	2.400	0.3000	$(s_2, +0.40)$	Low
c_5 Adherence	4.512	0.5640	$(s_5, -0.49)$	Med-High
c_6 Security	5.147	0.6433	$(s_5, +0.15)$	Med-High

of-Cards method have demonstrated its versatility for constructing interpretable fuzzy membership functions in a co-constructive analyst–decision-maker setting [59], and for modelling higher-order uncertainty through interval type-2 fuzzy sets [60]; our use of the method is limited to the classical SRF weight-elicitation protocol. In this synthetic study, the session was conducted by the research team acting in the role of the evaluating panel (Sect. 6.2.1). The protocol proceeds in three steps:

1. *Ranking with intensity:* The panel is given six cards, one per criterion $c_1, \ldots, c_6$, plus a set of blank cards. The cards are rank-ordered by judged importance for SDG 3 achievement in the target context, and blank cards are inserted between consecutive positions to express the intensity of the importance difference—the more blank cards, the larger the gap.
2. *Ratio anchoring:* The panel states the ratio z between the importance of the most and the least important criterion. Together with the card positions, this determines a ratio scale of non-normalised weights via the standard SRF computation [17].
3. *Pairwise interaction elicitation:* The SRF weights fix the singleton densities $g_i = m(\{i\})$ (their *relative* profile normalised so that $\sum_i g_i \leq 1$). The remaining mass is assigned to the pairwise interaction terms $m(\{i, j\})$, elicited only for those criterion pairs the panel identifies as interdependent. Rather than a holistic capacity-identification procedure over reference alternatives, we use targeted expert judgement on documented dependencies: for each candidate pair, a domain specialist estimates the coalition value $\mu(\{i, j\})$—the joint importance of holding both criteria together—from which the Möbius mass $m(\{i, j\}) = \mu(\{i, j\}) - g_i - g_j$ follows. Non-interacting pairs receive $m(\{i, j\}) = 0$ and need not be elicited. The densities and pairwise masses are then entered into the tool, which assembles the 2-additive capacity via Eq. (8) and checks the normalisation and monotonicity conditions of Lemma 1.

This two-stage design separates two cognitively distinct judgments—relative criterion importance (Steps 1–2, panel-wide) and localised pairwise interaction (Step 3, specialist-led)—and confines the Deck-of-Cards exercise to the six criteria, keeping the panel session short. The pairwise interaction terms are elicited from a smaller specialist input rather than the full panel, since judging *whether* and *how strongly* two specific criteria reinforce each other requires domain expertise that the broad multidisciplinary panel does not uniformly hold.

5.7.2 Elicited Densities and Interaction Structure

The panel's Deck-of-Cards session (Steps 1–2) produced the singleton densities reported in Table 8, with Rural Connectivity (c_4) and Telemedicine (c_1) ranked most important. The complete protocol is reproduced below for transparency.

Step 1—Ranking with intensity. The six criteria cards were ranked (bottom = least important, top = most important) as:

$$\{c_6, c_5\} \prec c_3 \prec c_2 \prec c_1 \prec c_4.$$

Data Security (c_6) and Treatment Adherence (c_5) were judged equally important and placed *ex aequo* in the lowest position, so the ranking comprises five distinct positions

rather than six. No blank cards were inserted between consecutive positions, denoting importance differences that the panel judged uniform and minimal along the ranking; the overall spread between the extremes is instead fixed by the ratio z elicited in Step 2.

Step 2—Ratio anchoring. The panel stated $z = g_4/g_6 = 2.0$ (the most important criterion is twice as important as the least). Assigning position values $p = (1, 1, 2, 3, 4, 5)$ to the ranked sequence $(c_6, c_5, c_3, c_2, c_1, c_4)$—the two *ex aequo* criteria sharing position 1, and each subsequent position incremented by one unit plus the number of blank cards inserted before it (here none), giving $p_{\max} = 5$—and applying the standard SRF formula [17]

$$e_i = 1 + (z - 1)\,\frac{p_i - 1}{p_{\max} - 1}, \qquad g_i = \frac{e_i}{\sum_j e_j},$$

yields the non-normalised weights $\mathbf{e} = (1.00, 1.00, 1.25, 1.50, 1.75, 2.00)$ and, after normalisation under the constraint $\sum_i g_i = 1 - \sum m(\{i, j\}) = 0.90$, the singleton densities $\mathbf{g} = (0.1059, 0.1059, 0.1324, 0.1588, 0.1853, 0.2118)$ reported in Table 8 (mapped back to the original criterion indices c_1–c_6). These values are displayed to four decimal places for readability; all computations reported in this section use the unrounded densities $g_i = 0.90\, e_i / \sum_j e_j$, which sum to exactly 0.90, and the released code constructs them from the SRF weights rather than from the rounded figures. For the pairwise interactions (Step 3), a telemedicine domain expert identified two dependencies as material for SDG 3 delivery in the target context, both rooted in the reliance of software-layer services on physical connectivity: Telemedicine depends on Rural Connectivity (c_1–c_4), and Treatment Adherence tracking likewise depends on it (c_4–c_5). All other criterion pairs were judged non-interacting and assigned $m(\{i, j\}) = 0$, following the independence default of the protocol of Sect. 4.3. The expert expressed each dependency as a positive synergy mass; the base ("medium") calibration places $m(\{c_1, c_4\}) = m(\{c_4, c_5\}) = 0.05$, i.e. coalition values $\mu(\{c_1, c_4\}) = 0.4471$ and $\mu(\{c_4, c_5\}) = 0.3676$. Table 8 lists the resulting capacity.

The value 0.05 is not arbitrary, and its construction is stated here explicitly. Three considerations fix it.

(i) Verbal-to-numerical anchoring. The expert did not state a number directly but graded each dependency on a three-point verbal scale—weak, medium, strong—anchored beforehand to the Möbius masses 0.03, 0.05 and 0.08. Both dependencies were graded *medium*. The anchors were chosen so that the total interaction mass $\sum m(\{i, j\}) = 2m$ remains a minority share of the capacity across the whole scale (6%, 10% and 16% respectively), ensuring that the measure stays predominantly determined by the panel-wide Deck-of-Cards densities rather than by the single specialist judgement that supplies the interaction terms.

(ii) Interpretation via the Shapley indices. For a 2-additive capacity the Shapley importance and interaction indices take the closed forms $\phi_i = g_i + \frac{1}{2}\sum_{j \neq i} m(\{i, j\})$ and $I_{ij} = m(\{i, j\})$ [12], which express the calibration in terms the expert can audit. Under the base calibration the importance of Rural Connectivity rises from its elicited density $g_4 = 0.2118$ to $\phi_4 = 0.2618$, an increase of 23.6%, while each of the two dependent criteria gains $m/2 = 0.025$. This is the reading the expert intended—connectivity is a prerequisite whose effective importance exceeds its stand-alone

Table 8 2-additive capacity specified for the case study (base calibration). Singleton densities constructed via the Deck-of-Cards (SRF) protocol; pairwise synergies assigned by the research team following domain-informed judgement (Sect. 6.2.1). For singletons, the Möbius mass equals the capacity value; for pairs, the capacity value $\mu(\{i, j\}) = g_i + g_j + m(\{i, j\})$ exceeds the sum of the individual densities, encoding the superadditive interaction

Parameter	Möbius mass m	Capacity μ
Singleton densities (Deck-of-Cards, $\sum g_i = 0.90$)		
g_1 Telemedicine adoption	0.1853	0.1853
g_2 Device access	0.1588	0.1588
g_3 Home accessibility	0.1324	0.1324
g_4 Rural connectivity	0.2118	0.2118
g_5 Treatment adherence	0.1059	0.1059
g_6 Data security	0.1059	0.1059
Pairwise interactions (expert judgement, $\sum m = 0.10$)		
$m(\{c_1, c_4\})$ Telemedicine–Rural	0.0500	0.4471
$m(\{c_4, c_5\})$ Rural–Adherence	0.0500	0.3676
$\mu(N) = \sum g_i + \sum m(\{i, j\})$	1.0000	

weight—without any single pair dominating the measure, since $I_{ij} = 0.05$ against importances in the range 0.11–0.26.

(iii) Ordinal consistency with the elicited ranking. The calibration is additionally bounded above by the requirement that the Shapley importances not contradict the ranking produced by the panel in Step 1. Since c_5 receives interaction mass and c_3 does not, increasing m eventually inverts the elicited order $c_3 \succ c_5$. Writing $\phi_5(m) = \phi_3(m)$ under the proportional rescaling $g_i' = g_i\,(1 - 2m)/\sum_j g_j$ and substituting the SRF weights $e_3 = 1.25$, $e_5 = 1.00$ with $\sum_j e_j = 8.5$ reduces to $\frac{1}{2}(1 - 2m) = 8.5\,m$, whose solution is the exact threshold $m^\star = 1/19 \approx 0.0526$: below it, the Shapley importance order reproduces the panel's Deck-of-Cards ranking exactly; above it, it does not. The base calibration $m = 0.05$ is therefore the largest anchor of the verbal scale that remains ordinally consistent with the panel elicitation, which is why it—and not the strong anchor—is adopted as the reference specification.

All Möbius masses are non-negative and sum to one, so by Lemma 1, the capacity is a well-formed normalised, monotone fuzzy measure. The two positive pairwise masses make the capacity *superadditive* on the interacting coalitions: for instance $\mu(\{c_1, c_4\}) = 0.4471 > g_1 + g_4 = 0.3971$, so Telemedicine and Rural Connectivity jointly contribute more than their individual densities suggest—the intended encoding of the infrastructure dependency.

5.7.3 Choquet Computation

The criterion values are sorted in increasing order,

$$c_4(0.300) \le c_3(0.460) \le c_2(0.513) \le c_1(0.555) \le c_5(0.564) \le c_6(0.643),$$

and the Choquet integral is evaluated over the resulting nested coalitions using the capacity of Table 8. Table 9 reports the full step-by-step computation; the same computation is reproduced for the weak and strong calibrations in Online Resource 1 (Tables S4–S6).

The retranslation gives the following:

$$\hat{\beta} = 0.4666 \times 8 = 3.733 \;\Rightarrow\; \hat{r} = \Delta_{\overline{S}}(3.733) = (s_4,\ -0.27).$$

The superadditive structure amplifies the influence of the worst-performing criterion: because c_4 enters first in the sorted sequence, its marginal coalition weight is $\Delta\mu_{(1)} = \Delta\mu_{c_4} = 0.3118 > g_4 = 0.2118$, so Rural Connectivity carries more weight in the aggregation than its density alone—explicitly encoding the systems-level dependency of the software-layer criteria on connectivity. The additive baseline (equal-weight 2t-WA), by contrast, yields $\hat{r} = (s_4,\ +0.05)$: the Choquet operator displaces the global evaluation downward by roughly 0.31 BLTS units relative to the additive case, recovering the structural severity of the c_4 deficit that a plain average absorbs into a neutral central reading.

5.7.4 Robustness to the Elicited Interaction Strength

Because the pairwise synergies are the one input drawn from specialist judgement rather than the panel-wide Deck-of-Cards session, we assess how sensitive the global evaluation is to their magnitude. The telemedicine expert bounded each dependency within a plausible range—weak, base, and strong—corresponding to Möbius masses of 0.03, 0.05, and 0.08 per interacting pair; in each case the singleton densities are rescaled proportionally so that $\mu(N) = 1$ is preserved. Table 10 reports the resulting global evaluations; the coalition-by-coalition computation behind each of the three calibrations is given in Online Resource 1 (Tables S4–S9).

Across the full plausible range of interaction strength, the global evaluation stays within $\hat{\beta} \in [3.644, 3.792]$—always in the lower half of *Medium* and always strictly below both additive baselines (density-weighted, 3.881; equal-weight, 4.048). The aggregation conclusion is therefore invariant to the exact synergy magnitude: stronger assumed dependency simply deepens the penalty attributed to the Rural Connectivity

Table 9 Step-by-step Choquet integral over the 2-additive capacity (base calibration)

Rank	$c_{(i)}$	$\hat{\beta}_{(i)}$	$\mu(A_{(i)})$	$\mu(A_{(i+1)})$	$\Delta\mu_i$	$\hat{\beta}_{(i)} \cdot \Delta\mu_i$	$A_{(i)}$
1	c_4 Rural	0.3000	1.0000	0.6882	0.3118	0.0935	$\{c_4, c_3, c_2, c_1, c_5, c_6\}$
2	c_3 Home	0.4600	0.6882	0.5559	0.1324	0.0609	$\{c_3, c_2, c_1, c_5, c_6\}$
3	c_2 Devices	0.5133	0.5559	0.3971	0.1588	0.0815	$\{c_2, c_1, c_5, c_6\}$
4	c_1 Telemedicine	0.5550	0.3971	0.2118	0.1853	0.1028	$\{c_1, c_5, c_6\}$
5	c_5 Adherence	0.5640	0.2118	0.1059	0.1059	0.0597	$\{c_5, c_6\}$
6	c_6 Security	0.6433	0.1059	0.0000	0.1059	0.0681	$\{c_6\}$
$\mathcal{C}_\mu(\mathbf{r})$						**0.4666**	

bottleneck, monotonically and with a clear domain interpretation. Only the identification of *which* criterion drives the result—and its attribution to a named dependency—is specific to the Choquet operator; the direction of the effect is robust to the elicitation.

One remark completes the picture. The strong calibration ($m = 0.08$) exceeds the ordinal-consistency threshold $m^\star = 1/19$ derived in Sect. 5.7: at that magnitude the Shapley importance of c_5 overtakes that of c_3 (0.1388 vs. 0.1235), inverting the order the panel stated in the Deck-of-Cards session. It is retained in Table 10 as a deliberate stress test—a specification the panel would not have endorsed—and the invariance reported above is correspondingly stronger than it appears, holding across a range of interaction strengths that extends beyond the ordinally admissible interval.

5.7.5 Sufficiency of the 2-Additive Representation

The restriction to pairwise interactions is a modelling assumption (Sect. 3.6.1) and is tested directly on the case study. Extending the base capacity with a third-order Möbius mass on the connectivity cluster $\{c_1, c_4, c_5\}$—the only triple for which a joint effect could plausibly be argued—and sweeping it across the range the domain expert regarded as conceivable, $[-0.03, +0.03]$, moves the global evaluation only within $\hat{\beta} \in [3.688, 3.777]$: a span of 0.089 BLTS units, smaller than the 0.148 displacement produced by the pairwise interactions themselves. The criterion ranking, the bottleneck attribution and the linguistic label are invariant throughout, and monotonicity holds at every value. The third-order term therefore modulates the magnitude of the effect but not the conclusion, which is what the 2-additive restriction assumes. This is evidence about this capacity and this profile rather than a general result. The full sweep is reported in Online Resource 1, Section S9.

5.8 Results and Comparison

5.8.1 Three-Level Tool Output

The tool produces results at three levels. At the **expert level**, the full 2-tuple assessment matrix (Online Resource 1, Table S3) provides per-expert per-criterion evaluations in the unified BLTS domain. At the **criterion level**, the per-criterion panel assessments

Table 10 Global evaluation as a function of the elicited pairwise synergy strength $m(\{i, j\})$ (with $\mu(N) = 1$ maintained). When varying the synergy mass, the singleton densities are rescaled proportionally: $g_i' = g_i \times (1 - \sum m'(\{i, j\}))/\sum_j g_j$, so that $\sum_i g_i' + \sum m'(\{i, j\}) = 1$ is preserved in every calibration

Calibration	$m(\{i, j\})$	$\hat{\beta}$	2-tuple	Label
Weak synergy	0.03	3.792	$(s_4, -0.21)$	Medium
Base	**0.05**	**3.733**	$(\mathbf{s_4}, \mathbf{-0.27})$	**Medium**
Strong synergy	0.08	3.644	$(s_4, -0.36)$	Medium
Density-weighted (additive, $m_{ij} = 0$)	0.00	3.881	$(s_4, -0.12)$	Medium
Equal-weight 2t-WA (reference)	—	4.048	$(s_4, +0.05)$	Medium

reported in Table 7 constitute the primary diagnostic output; ranked by performance, they yield the ordering $c_6 \succ c_5 \succ c_1 \succ c_2 \succ c_3 \succ c_4$. At the **global level**, the Choquet integral over the elicited 2-additive capacity yields $\hat{r} = (s_4, -0.27)$ (Table 9), with the robustness of this value to the elicited interaction strength established in Table 10.

Rural Connectivity (c_4) is the lowest-performing criterion across all scenarios and aggregation methods, constituting the primary structural bottleneck of the synthetic scenario. Whether connectivity plays this role in any real setting is not a question this case study can address.

Figures 5 and 6 show the results dashboard of the tool, presenting the three-level output for this case study.

5.8.2 Visualising the Choquet Effect

Figure 7 decomposes the shift from the density-weighted additive baseline ($\hat{\beta} = 3.881$) to the Choquet result ($\hat{\beta} = 3.733$). For each criterion it shows two quantities: the *additive* contribution $g_i^{\text{norm}} \times \hat{\beta}_i$ (the share of the global score attributable to criterion i under equal-density aggregation) and the *Choquet* contribution $\Delta\mu_i \times \hat{\beta}_i$ (its share under the 2-additive capacity). Rural Connectivity (c_4) is the only criterion whose Choquet contribution *exceeds* its additive one (0.748 vs. 0.565), because the superadditive capacity assigns it effective weight $\Delta\mu_4 = 0.3118$, substantially above its normalised density $g_4^{\text{norm}} = 0.2353$. Since c_4 is the lowest-performing criterion ($\hat{\beta}_4 = 2.40$), this amplified weight channels more of a below-average score into the global result, pulling it downward. The five remaining criteria all cede weight and contribute less under the Choquet operator; their combined reduction (-0.331) more than offsets the c_4 increase ($+0.184$), yielding the net displacement of -0.148.

5.8.3 Comparative Analysis Across Scenarios and Baselines

Table 11 reports the global evaluation under all aggregation methods. The criterion ranking $c_4 \prec c_3 \prec c_2 \prec c_1 \prec c_5 \prec c_6$ is invariant across all methods—Rural Connectivity is a robustly identified bottleneck regardless of aggregation choice. The methods diverge in their global score and in the degree to which they surface the structural severity of that bottleneck.

Across the plausible range of elicited interaction strength (Table 10) the global Choquet score lies in $\hat{\beta} \in [3.644, 3.792]$—*Medium* with a negative symbolic translation—strictly below the additive equal-weight and dispersion-weighted baselines ($\hat{\beta} \geq 4.048$), with Rural Connectivity (c_4) identified as the bottleneck in every case. Notably, every method in Table 11 produces the same base label *Medium* (s_4); the entire discrimination between aggregation operators is carried by the symbolic translation $\alpha \in [-0.35, +0.13]$, illustrating the importance of the continuous 2-tuple representation: a purely ordinal model would report identical outputs for all methods. The interpretive comparison of these methods—why the additive baselines absorb the c_4 deficit, why the OWA pessimistic operator reaches a similar range through a criterion-blind mechanism, and how the Choquet operator alone attributes the result to a specific criterion via the superadditive amplification $\Delta\mu_{(1)} = \Delta\mu_{c_4} = 0.3118 > g_4 = 0.2118$—is developed in Sect. 6.

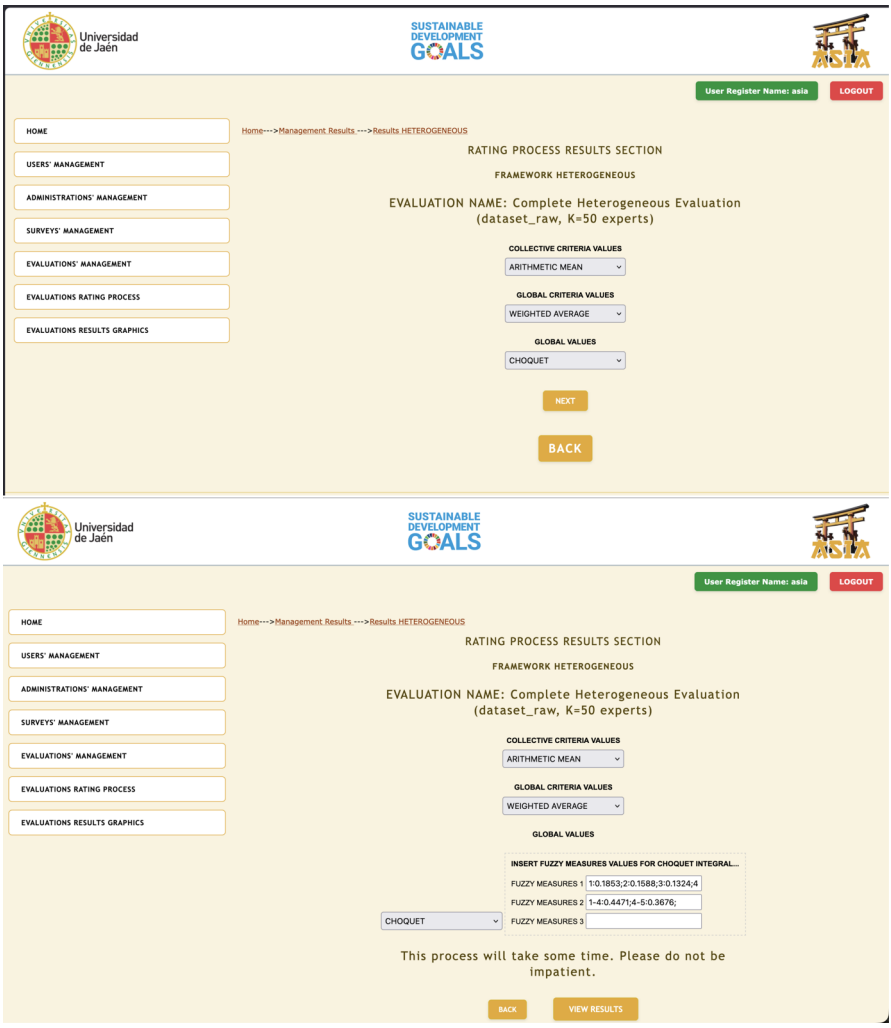
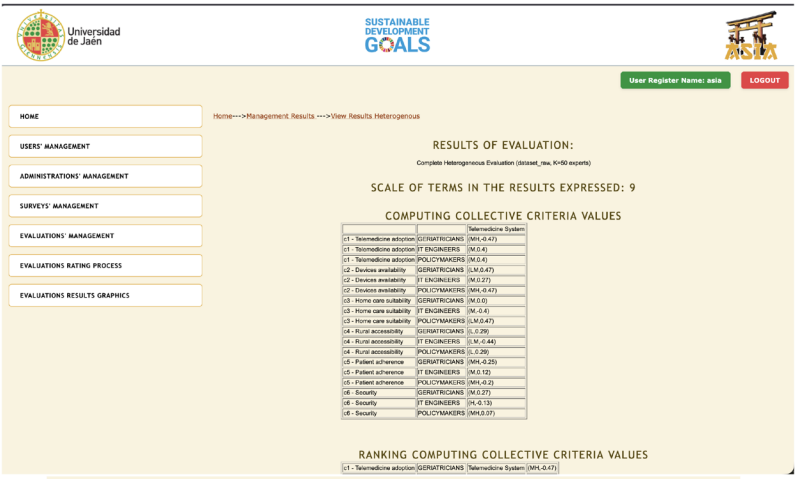


Fig. 5 Tool screen 3a: selection of the aggregation operators for the expert and criterion levels and Tool screen 3b: entry of the 2-additive capacity. FUZZY MEASURES 1 holds the singleton densities and FUZZY MEASURES 2 the pairwise coalition values, as reported in Table 8

5.8.4 Sensitivity to Interaction Strength

Figure 8 plots the global score $\hat{\beta}$ as a continuous function of the pairwise synergy mass $m(\{i, j\})$, extending the three discrete calibrations of Table 10 across the range $[0, 0.10]$ (with the singleton densities rescaled at each point so that $\mu(N) = 1$). The point $m_{ij} = 0$ recovers the density-weighted additive case ($\hat{\beta} = 3.881$); increasing synergy deepens the penalty on the bottleneck monotonically.



RANKING COMPUTING COLLECTIVE CRITERIA VALUES

c1 - Telemedicine adoption	GERIATRICIANS	Telemedicine System	(MH,-0.47)
c1 - Telemedicine adoption	IT ENGINEERS	Telemedicine System	(M,0.4)
c1 - Telemedicine adoption	POLICYMAKERS	Telemedicine System	(M,0.4)
c2 - Devices availability	GERIATRICIANS	Telemedicine System	(LM,0.47)
c2 - Devices availability	IT ENGINEERS	Telemedicine System	(M,0.27)
c2 - Devices availability	POLICYMAKERS	Telemedicine System	(MH,-0.47)
c3 - Home care suitability	GERIATRICIANS	Telemedicine System	(M,0.0)
c3 - Home care suitability	IT ENGINEERS	Telemedicine System	(M,-0.4)
c3 - Home care suitability	POLICYMAKERS	Telemedicine System	(LM,0.47)
c4 - Rural accessibility	GERIATRICIANS	Telemedicine System	(L,0.29)
c4 - Rural accessibility	IT ENGINEERS	Telemedicine System	(LM,-0.44)
c4 - Rural accessibility	POLICYMAKERS	Telemedicine System	(L,0.29)
c5 - Patient adherence	GERIATRICIANS	Telemedicine System	(MH,-0.25)
c5 - Patient adherence	IT ENGINEERS	Telemedicine System	(M,0.12)
c5 - Patient adherence	POLICYMAKERS	Telemedicine System	(MH,-0.2)
c6 - Security	GERIATRICIANS	Telemedicine System	(M,0.27)
c6 - Security	IT ENGINEERS	Telemedicine System	(H,-0.13)
c6 - Security	POLICYMAKERS	Telemedicine System	(MH,0.07)

COMPUTING GLOBAL CRITERIA VALUES

	Telemedicine System
c1 - Telemedicine adoption	(M,0.44)
c2 - Devices availability	(M,0.11)
c3 - Home care suitability	(M,-0.32)
c4 - Rural accessibility	(L,0.4)
c5 - Patient adherence	(MH,-0.49)
c6 - Security	(MH,0.15)

RANKING COMPUTING GLOBAL CRITERIA VALUES

c1 - Telemedicine adoption	Telemedicine System	(M,0.4)
c2 - Devices availability	Telemedicine System	(M,0.1)
c3 - Home care suitability	Telemedicine System	(M,-0.3)
c4 - Rural accessibility	Telemedicine System	(L,0.4)
c5 - Patient adherence	Telemedicine System	(MH,-0.5)
c6 - Security	Telemedicine System	(MH,0.2)

COMPUTING GLOBAL VALUES

Telemedicine System
(M,-0.27)

Fig. 6 Tool screen 3c: panel assessments disaggregated by professional group (geriatricians, IT engineers, policymakers) and Tool screen 3d: criterion-level results and the global Choquet evaluation. Rural Connectivity (c_4) is the only criterion falling below the *Medium* term (linguistic label *Low*), identifying it as the primary structural bottleneck of the synthetic scenario; the global evaluation is (s_4, -0.27)

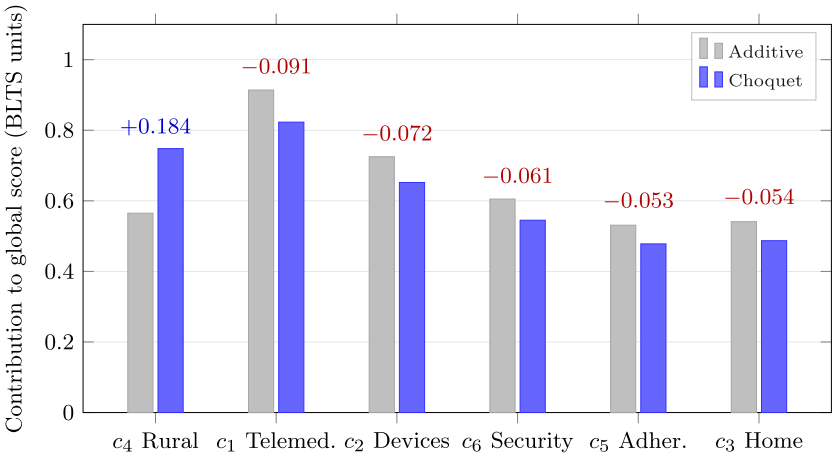


Fig. 7 Per-criterion decomposition of the Choquet displacement. Grey bars: additive (density-weighted) contribution $g_i^{\text{norm}} \times \hat{\beta}_i$. Blue bars: Choquet contribution $\Delta\mu_i \times \hat{\beta}_i$. Labels show the per-criterion shift. Rural Connectivity (c_4) is the only criterion that gains weight under the 2-additive capacity; since it is the worst-performing criterion, the amplified weight drives the global score downward by -0.148 BLTS units

5.8.5 Extended Sensitivity Analysis

The interaction strength analysed above is one of four inputs on which the global evaluation could in principle depend. The remaining three—the weights assigned to individual experts, uniform in the base analysis (Sect. 3.4); the granularity of the BLTS; and the elicited densities that parameterise the capacity—are perturbed here under a common protocol: 1000 replications per cell, a fixed seed, and the same reported quantities in each case, namely the distribution of $\hat{\beta}$, whether the reported linguistic label changes, and whether the qualitative conclusions survive. The detailed results are given in Online Resource 1, Section S5; Table 12 summarises all four axes.

Three points in the table warrant comment. **On Axis A**, the criterion ranking changes in 27% of replications at $\alpha = 5$, the most unequal panel examined. That figure overstates the instability: mean Kendall τ against the base ranking is 0.96, and the changes

Table 11 Global SDG 3 evaluation under all aggregation methods. Dispersion weights are computed as $w_i = \sigma_i / \sum_j \sigma_j$, where σ_i is the standard deviation of the unified expert assessments $\{\Delta_S^{-1}(s_i^k, \alpha_i^k)\}_{k=1}^K$ of criterion c_i, so that criteria with higher inter-expert disagreement receive larger weight

Method	$\hat{\beta}$	2-tuple	Label	Mechanism
Equal-weight 2t-WA	4.048	$(s_4, +0.05)$	Medium	Uniform weights
Dispersion-weighted 2t-WA	4.128	$(s_4, +0.13)$	Medium	Rewards disagreement
2-tuple OWA (pessimistic)	3.653	$(s_4, -0.35)$	Medium	Uniform pessimism
Density-weighted 2t-WA	3.881	$(s_4, -0.12)$	Medium	Additive ($m_{ij} = 0$)
Choquet (2-additive)	**3.733**	$(\mathbf{s_4}, \mathbf{-0.27})$	**Medium**	**Superadditive pairs**

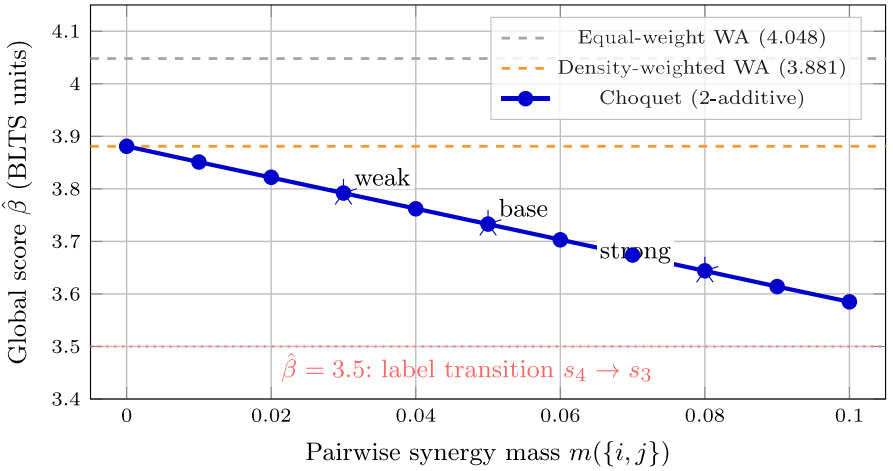


Fig. 8 Sensitivity of the global SDG 3 score to the elicited pairwise synergy mass $m(\{i, j\})$. The score remains in the (s_4, α^-) region—*Medium* with a negative symbolic translation—for the entire plausible range of interaction strength, confirming the robustness of the bottleneck identification. The label would only transition to *Low-Med* (s_3, α^+) if $\hat{\beta}$ fell below 3.5, which does not occur within the plotted range

are almost entirely a single adjacent transposition between c_1 and c_5, whose panel means differ by 0.07 BLTS units and are therefore not meaningfully distinguishable.

On Axis B, the invariance is exact—maximum deviation 0.0 across all six scales—which is the numerical counterpart of Proposition 1(vi). The axis also exhibits the limit of that property: the normalised value is invariant, but the *term* in which it is reported is not, since each scale partitions [0, 1] differently. Granularity determines the resolution of the linguistic report without affecting the aggregation.

On Axis C, the most consequential of the three, the label does change in 2.4% of replications under the widest perturbation. The two conclusions the case study actually rests on are nonetheless invariant: the Choquet score remains strictly below the additive baseline computed on the *same* perturbed densities in 100% of all 3000 replications, and the amplification condition $\Delta\mu_{c_4} > g_4$ holds in 100% of them. The displacement

Table 12 Summary of the four sensitivity axes. Each axis is perturbed independently over the range shown, with 1000 replications per cell for the stochastic axes (A, C). "Label stable" reports the percentage of replications in which the reported linguistic term is unchanged; "bottleneck stable" the percentage in which c_4 remains the lowest-ranked criterion

Axis	Range explored	$\hat{\beta}$ range	Label stable	Bottleneck stable
Interaction strength	$m \in [0.00,\ 0.10]$	3.585–3.881	100%	100%
A. Expert weights	Dirichlet, $\alpha \in \{5, 20, 100\}$	3.615–3.857	100%	100%
B. BLTS granularity	$g \in \{4, 6, 8, 10, 12, 16\}$	Invariant (normalised)	—	100%
C. Capacity densities	Dirichlet, $\alpha \in \{50, 200, 1000\}$	3.355–4.135	97.6%	100%

mechanism is a property of the capacity structure rather than of the particular density values elicited.

Taken together, the four axes support a bounded claim: the magnitude of the evaluation depends most on the elicited densities and least on the granularity of the scale, while the qualitative findings—the bottleneck and the direction of the Choquet displacement—are invariant across every perturbation examined.

5.8.6 Sensitivity and Robustness

The stability of the global evaluation across the full plausible range of elicited interaction strength (Table 10, Fig. 8)—always *Medium* with a negative symbolic translation—demonstrates that the aggregation conclusion is robust to the specific magnitude of the elicited synergies: regardless of how strongly the Telemedicine–Rural and Rural–Adherence dependencies are judged to interact, the framework consistently identifies Rural Connectivity as the primary bottleneck of the generated profile and places the global evaluation strictly below the equal-weight and dispersion-weighted baselines and below the centre of the *Medium* term. Varying the pairwise synergy from weak to strong shifts the global score within a range of only 0.15 BLTS units (3.644–3.792) without crossing a label boundary, making the aggregation result robust.

In a real deployment, this would suggest that targeted ICT infrastructure investment in dispersed rural areas is a prerequisite for the effectiveness of software-layer interventions such as telemedicine platforms or treatment adherence systems; however, since the present case study is fully synthetic, this observation serves only to illustrate the type of diagnostic insight that the framework can produce.

6 Discussion and Comparative Analysis

This section contextualises the results of Sect. 5 within the broader landscape of SDG aggregation methodologies, characterises computational complexity, and details the expert elicitation protocol underlying the case study.

6.1 Comparative Aggregation Analysis

The full comparison across aggregation methods and Choquet scenarios is reported in Table 11 of Sect. 5. We highlight here the key interpretive distinctions.

The equal-weight WA and dispersion-weighted WA both place the global score above the centre of the *Medium* term ($\hat{\beta} \geq 4.048$), absorbing the structural severity of the Rural Connectivity deficit into a neutral central reading. The OWA pessimistic operator reaches a score slightly below the Choquet result ($\hat{\beta} = 3.653$ vs. 3.733) but through a mechanism blind to criterion identity—it uniformly downweights high performers without encoding which criterion is the source of underperformance. Neither weighted-average operator provides a rationale grounded in the domain structure.

The 2-additive Choquet operator, across the full plausible range of specified interaction strength, consistently displaces the global score below the additive equal-weight and dispersion-weighted baselines ($\hat{\beta} \in [3.644, 3.792]$ vs. ≥ 4.048), with Rural Connectivity as the identified bottleneck. This is the key robustness finding: the aggregation

conclusion is invariant to the specified synergy magnitude, while the Choquet framework provides the formal language to express *which* criteria interact and to attribute the global displacement to a named dependency—something no additive baseline can do. Unlike the OWA operator, which reaches a comparable score by uniform pessimism, the Choquet operator localises the effect to the connectivity-dependent coalitions.

A further point of comparison concerns the DCM-based Choquet tradition in MCDA [13–15], which represents the methodological state of the art in capacity construction. These approaches infer a general (or k-additive) capacity from holistic preference statements of a decision maker, typically via linear programming over Deck-of-Cards-derived scales. The present framework shares their k-additive representation but constructs the capacity differently: rather than solving an inverse problem from holistic alternative comparisons, it elicits the singleton densities from a broad panel via Deck-of-Cards and the pairwise interaction terms from targeted specialist judgement on documented dependencies. Recent work by García-Zamora et al. [59] and Dutta et al. [60] has extended the Deck-of-Cards method to construct interpretable fuzzy membership functions and interval type-2 fuzzy sets, respectively, broadening its applicability beyond weight elicitation; these extensions are complementary to the present use and could enrich the framework's elicitation layer in future work. DCM-based identification is designed for a single decision maker or a small homogeneous group evaluating alternatives on a common representational domain; none of the cited works addresses the unification of heterogeneous multi-granular linguistic assessments from a large multidisciplinary panel, nor produces the exact linguistic outputs required for practitioner-facing SDG reporting. The proposed framework occupies this complementary niche, while adopting the Deck-of-Cards protocol precisely where the two traditions meet: the elicitation of the densities that parameterise the fuzzy measure (Sect. 5.7).

6.2 Expert Elicitation Protocol

6.2.1 Simulation Protocol

The panel is fully synthetic: no real experts participated in the assessment phase, and the case study makes no empirical claim about any real territory or deployment. The panel was generated via a seeded structured probabilistic protocol designed to replicate the representational heterogeneity and inter-group divergence patterns of real multidisciplinary evaluation panels. For each criterion c_i and professional group, a latent quality level was drawn from a truncated normal distribution on $[0, g]$ with group-specific location and dispersion parameters, and then expressed in the format prescribed for that criterion (Sect. 5): rounded to an integer score $v \in \{0, \ldots, 10\}$ for the numeric criteria (c_4, c_5), or snapped to the nearest term of the group's assigned scale for the linguistic criteria. All random draws are governed by a single fixed seed, so re-running the released generator reproduces the dataset—and hence every table in Sect. 5—exactly. Table 13 reports the empirical group-level characterisation of the generated dataset. Three designed properties are visible: coarse-granularity dispersion, where the S^3 scales assigned to the low-knowledge group of each criterion induce the

largest within-group variability (c_3 for IT engineers, $\sigma = 2.21$; c_6 for geriatricians, $\sigma = 2.37$); inter-group divergence on contested criteria (c_2: G 3.47 vs. P 4.53; c_6: G 4.27 vs. IT 5.87); and a uniformly depressed profile on c_4 (Rural Connectivity) across all three groups—the structural bottleneck that the aggregation analysis of Sect. 5 is designed to stress-test.

The complete generated dataset is reported in full in Online Resource 1 and released in machine-readable form together with the seeded generator itself, so that both the dataset and every result in Sect. 5—the unification, all three aggregation levels, the Choquet computations, and the sensitivity analysis—are exactly reproducible from the released code. Consistent with the fully synthetic design, the Deck-of-Cards session of Sect. 5.7 was conducted by the research team acting in the role of the evaluating panel.

6.2.2 Scope and Interpretation of the Synthetic Validation

Because both the assessment matrix and the density elicitation are synthetic, the purpose and the limits of what the case study establishes must be stated precisely. The study is a *controlled computational validation*: it verifies the mathematical robustness, the heterogeneity-handling capacity, and the discriminative behaviour of the framework and the tool under conditions that are fully specified and exactly reproducible. It makes no empirical claim about any real territory, population, or deployment, and it does not—and cannot—establish external validity. Establishing the latter requires real elicitation panels and is deferred to future work.

Within these limits, the study is designed so that its conclusions do not reduce to a tautology of the generator. Three points support this. First, the design parameters of the generator are not chosen to force a particular aggregation outcome but to reproduce two representational phenomena that are documented in the multidisciplinary-panel literature and that the framework must be able to absorb: cross-format heterogeneity (numeric versus multi-granular linguistic assessment across criteria) and inter-group divergence driven by differential domain knowledge (finer granularity where a group is expert, coarser where it is not). The group-level statistics of Table 13 show that these phenomena are present in the generated data—for instance, the largest within-group dispersion arises exactly on the coarse-scale criteria assigned to each low-knowledge group (c_3 for IT engineers, c_6 for geriatricians), and the inter-group gaps on the contested criteria (c_2, c_6) are of the magnitude reported in real panels. Second, the discriminative result the study reports is not the identification of the bottleneck *per se*—

Table 13 Empirical group-level characterisation of the released synthetic dataset: mean (standard deviation) of the unified assessments $\beta \in [0, 8]$ per professional group and criterion, computed from the unified dataset (Online Resource 1, Table S3)

Group	c_1	c_2	c_3	c_4	c_5	c_6
G—Geriatricians	4.53 (0.84)	3.47 (1.19)	4.00 (1.59)	2.29 (0.73)	4.75 (0.98)	4.27 (2.37)
IT—Engineers	4.40 (1.05)	4.27 (0.93)	3.60 (2.21)	2.56 (1.18)	4.12 (1.38)	5.87 (1.18)
P—Policymakers	4.40 (1.72)	4.53 (1.60)	3.47 (1.60)	2.29 (1.17)	4.80 (1.39)	5.07 (1.28)

any monotone aggregator ranks c_4 last, as Table 11 confirms—but the *degree* to which each operator recovers the structural severity of that bottleneck, and only the Choquet operator attributes the displacement to a named criterion through the superadditive amplification $\Delta\mu_{(1)} = \Delta\mu_{c_4} = 0.3118 > g_4 = 0.2118$. This behavioural distinction between operators is a property of the aggregation machinery, not of the generated data: it would hold for any dataset exhibiting a single depressed criterion, and the synthetic scenario merely provides a transparent instance in which it can be inspected step by step. Third, because the densities are elicited independently of the assessment matrix and the pairwise synergies are varied across their full plausible range, the robustness finding—that the aggregation conclusion is invariant to the elicited interaction strength (Sect. 6)—is a statement about the stability of the operator across elicitation choices, which a generator tuned to a single outcome could not produce.

It should be noted that the identification of c_4 as the bottleneck is *planted* in the generator parameters ($\mu_{c_4} = 2.40$ vs. 3.4–5.7 for the remaining criteria) and reinforced by the placement of both synergies on pairs containing c_4. The case study therefore demonstrates that the Choquet operator *recovers* a known structural deficit and *quantifies* its amplification, rather than *discovering* an unexpected one. The more informative evidence comes from Scenario 3 of Table 6, where moving the synergies to non-bottleneck criteria reverses the sign of the gap—a result that is not predetermined by the generator design.

Validation on real elicitation panels—for both the assessment matrix and the density elicitation—constitutes the most immediate direction for future work.

6.3 Comparison with Distance-Based Ranking Procedures

Section 2.1 argued that TOPSIS and the proposed framework address different decision problems; the argument is examined here on identical data. Since a closeness coefficient is undefined for a single evaluated unit, the comparison requires a set of alternatives: we generate 20 synthetic territories with the same seeded heterogeneous-panel protocol used throughout ($K = 50$ experts each), unify every panel with the same 2-tuple machinery, and apply four procedures to the *same* unified criterion matrix with the *same* elicited weights—the additive weighted average, classical TOPSIS with vector normalisation, TOPSIS on the normalised 2-tuple values, and the proposed 2-additive Choquet framework.

The Rankings Agree Spearman correlation between the Choquet framework and TOPSIS is $\rho = 0.9895$ (Kendall $\tau = 0.9474$), with 5 discordant pairs out of 190; the two TOPSIS variants rank identically ($\rho = 1.0000$). This is the expected result and we report it as such: on data where the criterion profile is the dominant signal, any reasonable aggregation procedure will order the alternatives similarly, and we make no claim of superior ranking accuracy.

What the Procedures Return Differs TOPSIS returns a dimensionless coefficient—the top-ranked territory scores 0.8042—which conveys position within the comparison set and nothing else. The proposed framework returns $(s_6, -0.05)$, *High*, a term on the scale the panel itself used, together with the identification of c_2 as the criterion

driving the shortfall. Neither the term nor the attribution is available from a closeness coefficient.

The Coefficient Is Set-Dependent; The Evaluation Is Not Removing the worst-performing territory and recomputing changes every TOPSIS coefficient, by up to 0.0349, and alters the relative position of 5 of the remaining 19 alternatives. The Choquet evaluation of those 19 is unchanged to machine precision (0.0000), because it never referred to the comparison set. For SDG assessment this distinction matters: a territory's reported performance should not depend on which other territories happen to be in the sample.

The conclusion is not that TOPSIS ranks worse, but that it answers a different question: where the task is to select among alternatives it is appropriate and the present framework offers no advantage, and where the task is to evaluate one unit on an interpretable scale and identify what limits it, the closeness coefficient does not provide what is needed. Per-alternative scores and ranks under all four procedures are given in Online Resource 1, Section S7.

6.4 Computational Complexity

The computational cost of the proposed framework is analysed across its four stages for a panel of K experts, n criteria, and BLTS granularity g.

Stages (i)–(iii): Heterogeneous Transformation Normalisation, fuzzy set mapping, and 2-tuple conversion each operate in $O(Kn)$. The cumulative cost of the transformation pipeline is $O(Kng)$, dominated by the fuzzy set construction step. For typical case study dimensions ($K = 50$, $n = 6$, $g = 8$), this is negligible.

Capacity Assembly The 2-additive capacity is specified directly by its Möbius masses—n singleton densities and the elicited pairwise terms (at most $\binom{n}{2}$, typically far fewer). No root-finding or optimisation is involved: assembling the masses and checking the monotonicity conditions of Lemma 1 is $O(n^2)$.

Fuzzy Measure Evaluation Evaluating $\mu(A)$ from Eq. 8 requires summing the singleton and pairwise masses contained in A, which is $O(|A|^2)$. The Choquet integral requires evaluating n nested coalitions after sorting, yielding a total cost of $O(n^3)$ in the worst case—negligible at the criterion counts relevant here. This is the dominant term in the aggregation stage.

Overall Complexity The complete pipeline runs in $O(Kng + n^3)$ time. For the dimensions typical of SDG multi-expert assessment ($K \leq 100$, $n \leq 20$, $g \leq 12$), this is computationally trivial: every stage is closed-form and no numerical optimisation is required. The framework is therefore suitable for interactive, real-time decision support at the panel sizes relevant to SDG evaluation.

Dimensional Scalability The primary scalability constraint of a general Choquet integral is the 2^n size of an unrestricted capacity. The 2-additive representation resolves this: the capacity is fully determined by the singleton densities and pairwise masses regardless of the number of subsets, enabling exact Choquet computation in polynomial time rather than $O(n \cdot 2^n)$. Where third- or higher-order interactions are documented, the same Möbius machinery extends to k-additive capacities [12] with $k \geq 3$, preserving the theoretical properties established in Sect. 3.

6.4.1 Measured Runtime

The complexity analysis above is verified empirically by timing the complete pipeline over 20 runs per configuration; the full grid is reported in Online Resource 1, Section S10. Absolute timings are machine-dependent and the quantity of interest is the scaling. Runtime is effectively flat in the panel size, since the expert-level aggregation is a single vectorised weighted mean and fixed overhead dominates at realistic panel sizes. It grows moderately in the number of criteria—a five-fold increase from $n = 6$ to $n = 30$ costs a factor of 5.3—consistent with the polynomial behaviour the 2-additive representation guarantees and far from the exponential growth an unrestricted capacity would produce. Granularity has no measurable effect, since g enters only as a scaling constant. The largest configuration examined, $K = 1000$ experts on $n = 30$ criteria at $g = 16$, completes in 0.588 ms, so the aggregation is not a practical constraint at any panel size a real elicitation exercise would produce.

7 Limitations and Scope

Several boundary conditions on the framework's applicability merit explicit acknowledgement. In particular, (i) the entire validation is synthetic, with no real experts involved; (ii) the quality of the aggregation outcome depends directly on the quality of the elicited densities and interaction terms; (iii) although the cost of the aggregation is polynomial and negligible at the panel and criteria counts relevant here, scaling to a substantially larger criteria set under higher-order capacities would require renewed attention to both computation and elicitation effort; and (iv) the tool's usability has not been formally evaluated. These points are developed below.

Regarding (iv), the computational and the human dimensions of the tool should be distinguished. The former is now measured (Sect. 6.4.1): computation is not a practical constraint at any realistic scale. The latter is not, and we make no claim about it. A usability evaluation requires a sample of practitioners from the professional groups the framework targets, a task protocol, and institutional ethics approval; conducting it properly is a study in its own right, and conducting it on a convenience sample of researchers already familiar with the formalism would produce a figure that could not support the conclusion it appeared to license. We therefore leave the System Usability Scale evaluation as declared future work (Sect. 8).

The 2-additive capacity captures interactions up to pairwise order only. Contexts with genuine higher-order effects—where three or more criteria interact jointly in a way not reducible to their pairwise relations—would require a k-additive capacity

with $k \geq 3$ [12]. For the case study, this restriction was tested rather than assumed: extending the capacity with a third-order mass on the connectivity cluster leaves the ranking and the global label unchanged across its plausible range (Sect. 5.7.5). That test speaks to this profile, not to domains in general. The same Möbius representation supports the extension directly, at the cost of eliciting the higher-order terms. Relatedly, the pairwise interaction terms are the one input drawn from targeted specialist judgement rather than the panel-wide Deck-of-Cards session. While the sensitivity analysis (Sect. 5.8.5) shows the aggregation conclusion to be robust to their magnitude, a systematic protocol for eliciting them—for instance via DCM-based ordinal regression over holistic preference statements [13, 14]—would strengthen the reproducibility of the interaction structure and is left for future work.

The elicitation of fuzzy densities $\mathbf{g}$ requires a structured prior elicitation phase. In general deployments, the quality of the density elicitation is a key determinant of aggregation quality, and pilot testing is recommended when applying the framework to a new domain.

The interval-valued input domain ($T_{I\overline{S}}$, Sect. 3.2) is supported by the framework and the tool, and is illustrated analytically in Sect. 5, but is not exercised end-to-end in the case study, which uses only numeric and multi-granular linguistic inputs. End-to-end validation with interval-valued assessments is left for future work with real panels.

The factorial simulation design (Sect. 5.2) covers the core interaction patterns (additive, synergy, redundancy) and their interaction with the position of the worst-performing criterion. The additional scenarios of Sect. 5.2.4 extend this to low-density bottlenecks and non-bottleneck interactions, confirming that the displacement mechanism generalises beyond the single-pair, symmetric-magnitude configuration. Scenarios with multiple simultaneous interactions across three or more pairs, or with negative (redundancy) masses of varying magnitude, remain untested and are natural extensions.

Finally, the framework aggregates the panel directly, without an intermediate Consensus Reaching Process (CRP)—a component common in group decision-making with 2-tuple linguistic information, in which experts iteratively revise their assessments to increase agreement before aggregation [3]. Here, inter-expert divergence is instead preserved and surfaced diagnostically at the expert and criterion levels, which is appropriate for the multidisciplinary setting targeted, where divergence is informative rather than noise to be reconciled. Integrating an optional CRP stage—allowing convergence to be pursued where it is warranted while retaining the divergence-preserving behaviour by default—is a natural extension left for future work.

8 Conclusions and Future Work

This paper addressed two interrelated limitations of existing SDG assessment frameworks: their inability to natively accommodate heterogeneous expert representations without additional numerical loss, and the absence of operational tools that make principled non-additive aggregation directly accessible to non-specialist practitioners. The proposed Heterogeneous Fuzzy Linguistic Model (HFLM) addresses both

through a three-level pipeline—expert, criterion, global—grounded in the 2-tuple linguistic representation and implemented in a decision support tool. Its computational performance is characterised in Sect. 6.4.1; its usability with practitioners has not been formally evaluated, and a study with the System Usability Scale is left for future work.

Computational verification against analytical baselines confirmed that the implementation correctly realises the equations of the model, including the algebraic identity by which the Choquet integral degenerates to the density-weighted average under an additive capacity. A controlled simulation experiment under a 3×2 factorial design then confirmed that (i) the Choquet–additive gap reverses sign under redundancy, (ii) the displacement is five times larger when the worst criterion participates in the synergy pair, and (iii) the variability of both the gap and the global score decreases with panel size once the ground truth is held fixed. Additional scenarios confirmed that the displacement mechanism generalises to low-density bottlenecks and that the gap reverses sign when the synergies are moved to non-bottleneck criterion pairs. A fully synthetic illustrative application—a simulated 50-expert panel on IoT-enabled healthcare for ageing populations (SDG 3), with the complete dataset published and all results exactly reproducible from it—demonstrated the discriminative capacity of the framework and the tool under controlled conditions. No real experts participated, and the case study does not support empirical conclusions about any real deployment. The global aggregation used a 2-additive capacity combining Deck-of-Cards densities with two pairwise synergies specified following the SRF protocol (Telemedicine–Rural and Rural–Adherence), yielding a global evaluation of $(s_4, -0.27)$—*Medium* with a negative symbolic translation—consistently below the centred *Medium* assessments yielded by the equal-weight and dispersion-weighted baselines ($\hat{\beta} \geq 4.048$). The sensitivity analysis confirmed that the global label is stable across the full plausible range of the specified interaction strength ($\hat{\beta} \in [3.644, 3.792]$), and that the criterion ranking $c_4 \prec c_3 \prec c_2 \prec c_1 \prec c_5 \prec c_6$ is invariant across all methods, making Rural Connectivity a robustly identified structural bottleneck *of the generated profile*. No empirical claim about SDG 3 performance in any real territory follows from this, and none is made.

Although illustrated in the SDG 3 context, the framework and tool are designed as a domain-agnostic architecture applicable to any MCDM context where expert panels are multidisciplinary and heterogeneous, criteria may interact non-additively, and practitioners require interpretable multi-level outputs. Candidate domains include supplier evaluation in heterogeneous procurement panels, clinical trial outcome assessment with multidisciplinary review boards, and policy impact evaluation across interdependent governance dimensions.

Several directions for future research are identified. First, validation of the tool with real expert panels—beyond the fully synthetic simulation used here—is the most immediate priority for establishing practical usability and external validity. Second, extending the expert-level aggregation to group-stratified OWA operators would allow differential treatment of expert subgroups (e.g. clinicians vs. engineers) within the same evaluation session. Third, extending the fuzzy measure beyond the pairwise (2-additive) representation to higher-order k-additive capacities elicited via DCM-based ordinal regression [13–15] would allow the framework to capture joint interactions among three or more criteria, and would replace the targeted expert judgement of the

pairwise terms with a systematic identification protocol, for smaller specialist panels able to sustain the richer elicitation. Fourth, incorporating Hesitant Fuzzy Linguistic Term Sets (HFLTS) [25] at the elicitation stage would allow experts to express intra-personal uncertainty between adjacent labels. Finally, application to additional SDG domains with larger and more geographically distributed panels would provide broader empirical grounding for the framework's generalisability.

Supplementary Information

Additional material is provided as Online Resource 1 (ESM_1.pdf), containing: Section S1, the 17 Sustainable Development Goals of the UN 2030 Agenda (Table S1); Section S2, the complete synthetic dataset in both native and unified BLTS representations (K = 50 experts, n = 6 criteria; Tables S2–S3), also available in machine-readable form in the replication package; Section S3, worked step-by-step Choquet integral calculations for the base, weak and strong calibrations of the robustness analysis (Tables S4–S6); Section S4, the proof of Proposition 1; Section S5, detailed results of the extended sensitivity analysis (Tables S7–S9); Section S6, detailed results of the strengthened validation (Tables S10–S13); Section S7, full results of the comparison with distance-based ranking procedures (Table S14); Section S8, the elicitation protocol for the interaction structure; Section S9, the full third-order sweep (Table S15); and Section S10, the runtime benchmark (Table S16).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43069-026-00715-1.

Author Contribution All authors contributed to the study conception and design. Model formulation, computational verification and the simulation experiment were performed collaboratively by F.J.M., S.C. and M.E. The decision support tool was designed and implemented by S.C. The main manuscript text was written by F.J.M., and all authors reviewed and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability The complete synthetic dataset (heterogeneous assessment matrix and unified BLTS representation, in machine-readable form), the seeded generator that produces it, and the analysis scripts reproducing every table and figure of the case study are publicly available at http://serezade.ujaen.es:8030/fjmimbre/f-sdg-2t-cho-doc

Code Availability The seeded dataset generator (`generate_dataset.py`), the case-study analysis code (`reproduce_results.py`, Python standard library only), and the simulation experiment (`simulation_experiment.py`, requiring NumPy) are included in the repository above, together with the full source code of the decision support tool (Java, JDK 17), its build configuration and deployment instructions for a standard servlet container. The methodology the tool implements is specified in full in Sect. 3, so that every result reported in Sect. 5 can be reproduced either through the tool or independently from the released Python scripts.

Declarations

Ethics Approval Not applicable. This study involves no human participants or animals; the expert panel used in the case study is entirely synthetic and generated computationally.

Conflict of Interest The authors declare no competing interests.

References

1. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning–I. Inf Sci 8(3):199–249. https://doi.org/10.1016/0020-0255(75)90036-5
2. Herrera F, Martínez L (2000) A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 8(6):746–752. https://doi.org/10.1109/91.890332
3. Herrera F, Martínez L (2001) A model based on linguistic 2-tuples for dealing with multigranular hierarchical linguistic contexts in multi-expert decision-making. IEEE Trans Syst Man Cybern B Cybern 31(2):227–234. https://doi.org/10.1109/3477.915345
4. Espinilla M, Liu J, Martínez L (2011) An extended hierarchical linguistic model for decision-making problems. Comput Intell 27(3):489–512. https://doi.org/10.1111/j.1467-8640.2011.00385.x
5. United Nations Statistics Division (2025) Global indicator framework for the sustainable development goals and targets of the 2030 agenda for sustainable development (2025 refinement). Technical report. https://unstats.un.org/sdgs/indicators/Global%20Indicator-Framework-after-2025-review-English.pdf
6. ONU (2020) Transforming our World: The 2030 Agenda for Sustainable Development. https://www.unfpa.org/resources/transforming-our-world-2030-agenda-sustainable-development
7. Triantaphyllou E (2000) Multi-criteria decision making methods: a comparative study. Applied Optimization, vol 44. Springer, Boston, MA
8. Sachs J, Schmidt-Traub G, Kroll C, Lafortune G, Fuller G (2019) Sustainable Development Report 2019. Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN), New York
9. Kahraman C (2018) Fuzzy multi-criteria decision-making: theory and applications. Springer, Cham
10. Herrera F, Martínez L, Sánchez PJ (2005) Managing non-homogeneous information in group decision making. Eur J Oper Res 166(1):115–132. https://doi.org/10.1016/j.ejor.2003.11.031
11. Grabisch M (1997) *k*-order additive discrete fuzzy measures and their representation. Fuzzy Sets Syst 92(2):167–189. https://doi.org/10.1016/S0165-0114(97)00168-1
12. Grabisch M, Marichal J-L, Mesiar R, Pap E (2009) Aggregation functions. Cambridge University Press, Cambridge
13. Bottero M, Ferretti V, Figueira JR, Greco S, Roy B (2018) On the Choquet multiple criteria preference aggregation model: theoretical and practical insights from a real-world application. Eur J Oper Res 271(1):120–140. https://doi.org/10.1016/j.ejor.2018.04.022
14. Barbati M, Greco S, Lami IM (2024) The Deck-of-cards-based Ordinal Regression method and its application for the development of an ecovillage. Eur J Oper Res 319(3):845–861. https://doi.org/10.1016/j.ejor.2024.07.010
15. Corrente S, Greco S, Zappalà S (2025) Deck of Cards method for hierarchical, robust and stochastic ordinal regression. Eur J Oper Res 327(3):937–956. https://doi.org/10.1016/j.ejor.2025.05.025
16. Espinilla M, Andrés Calle R, Martínez FJ, Martínez L (2013) A 360-degree performance appraisal model dealing with heterogeneous information and dependent criteria. Inf Sci 222:459–471. https://doi.org/10.1016/j.ins.2012.08.015
17. Figueira J, Roy B (2002) Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure. Eur J Oper Res 139(2):317–326. https://doi.org/10.1016/S0377-2217(01)00370-8
18. Corrente S, Figueira JR, Greco S (2021) Pairwise comparison tables within the deck of cards method in multiple criteria decision aiding. Eur J Oper Res 291(2):738–756. https://doi.org/10.1016/j.ejor.2020.09.036
19. Díaz-Sarachaga JM, Jato-Espino D (2018) A systematic literature review on the application of multi-criteria decision making methods in sustainable development. Sustain Dev 26(6):760–777

20. Allen C, Metternicht G, Wiedmann T (2021) Data-driven approaches for monitoring the sustainable development goals. Nature Sustainability 4:906–914
21. Sachs JD, Lafortune G, Woolston G, Yoshida S-Y (2022) Sustainable development report 2022. Cambridge University Press
22. Stiglitz JE, Sen A, Fitoussi J-P (2009) Report by the commission on the measurement of economic performance and social progress. Technical report, Commission on the Measurement of Economic Performance and Social Progress, Paris
23. Mordeson JN (2020) Fuzzy logic applied to sustainable development goals. Symmetry 12(1):87. https://doi.org/10.3390/sym12010087
24. Pradhan P, Costa L, Rybski D, Lucht W, Kropp J (2017) A systematic study of sustainable development goal interactions. Earth's Future 5(11):1169–1179. https://doi.org/10.1002/2017EF000632
25. Rodríguez RM, Martínez L, Herrera F (2012) Hesitant fuzzy linguistic term sets for decision making. IEEE Trans Fuzzy Syst 20(1):109–119. https://doi.org/10.1109/TFUZZ.2011.2170076
26. Rodríguez RM, Martínez L, Herrera F (2019) Multi-granular linguistic decision making: a review. Information Fusion 45:10–29
27. Kumar K, Chen S-M (2022) Group decision making based on weighted distance measure of linguistic intuitionistic fuzzy sets and the TOPSIS method. Inf Sci 611:660–676. https://doi.org/10.1016/j.ins.2022.08.088
28. Khan SZ, Akhtar Y, Salameh WMM, Karabasevic D, Stanujkic D (2026) A novel integrated group decision-making framework for assessing green supply chain strategies under complex uncertainty. Systems 14(4):418. https://doi.org/10.3390/systems14040418
29. Khan SZ, Akhtar Y, Salameh WMM, Syam MI (2026) A soft computing TOPSIS framework with fractional orthopair fuzzy sine aggregation for robust model evaluation in data science. Sci Rep 16:21637. https://doi.org/10.1038/s41598-026-49035-8
30. Zhang X, Jia P, Khan SZ, Syam MI, Rahim M, Khalifa HAE-W (2025) Fractional orthopair fuzzy decision framework for sustainable water resource management in urban areas. Sci Rep 15:42888. https://doi.org/10.1038/s41598-025-27126-2
31. Lourenzutti R, Krohling RA (2016) A generalized TOPSIS method for group decision making with heterogeneous information in a dynamic environment. Inf Sci 330:1–18. https://doi.org/10.1016/j.ins.2015.10.005
32. Lourenzutti R, Krohling RA, Reformat MZ (2017) Choquet based TOPSIS and TODIM for dynamic and heterogeneous decision making with criteria interaction. Inf Sci 408:41–69. https://doi.org/10.1016/j.ins.2017.04.037
33. Pang Q, Wang H, Xu Z (2016) Probabilistic linguistic term sets in multi-attribute group decision making. Inf Sci 369:128–143. https://doi.org/10.1016/j.ins.2016.06.021
34. Zadeh L (1999) From computing with numbers to computing with words - from manipulation of measurements to manipulation of perceptions. IEEE Transactions on Circuits and Systems I Fundamental Theory and Applications 46(1):105–119
35. Martínez L, Ruan D, Herrera F (2010) Computing with words in decision support systems: an overview on models and applications. Int J Comput Intell Syst 3(4):382–395
36. Mendel J, Zadeh L, Trillas E, Yager R, Lawry J, Hagras H, Guadarrama S (2010) What computing with words means to me: discussion forum. IEEE Comput Intell Mag 5(1):20–26
37. Yager R (2004) On the retranslation process in Zadeh's paradigm of computing with words. Systems Man and Cybernetics Part B Cybernetics IEEE Transactions on 34(2):1184–1195. https://doi.org/10.1109/TSMCB.2003.821866
38. Herrera F, Martínez L, Sánchez PJ (2008) A model for managing heterogeneous linguistic information in a multicriteria decision making context. Internat J Uncertain Fuzziness Knowledge-Based Systems 16(supp01):1–17
39. Li D-F, Huang Z-G, Chen G-H (2010) A systematic approach to heterogeneous multi-attribute group decision making. Computers Industrial Engineering 59(2):561–572
40. Delgado M, Herrera F, Herrera-Viedma E, Martínez L (1998) Combining numerical and linguistic information in group decision making. Inf Sci 107(1–4):177–194. https://doi.org/10.1016/S0020-0255(97)10044-5
41. Li H, Yang J, Xiang Z (2022) A fuzzy linguistic multi-criteria decision-making approach to assess emergency suppliers. Sustainability 14(20):13114. https://doi.org/10.3390/su142013114
42. Choquet G (1953) Theory of capacities. Annales de l'Institut Fourier 5:131–295. https://doi.org/10.5802/aif.53

43. Sugeno M (1974) Theory of Fuzzy Integrals and Its Applications. Tokyo Institute of Technology, Tokyo. Doctoral dissertation
44. Ramsawak RA, Behera DK, Nayak SK (2025) Digital health solutions and its usage by ageing populations in Asia: A bibliometric analysis and review. SAGE Open 15(2). https://doi.org/10.1177/21582440251396495
45. Tsai YI, Beh J, Ganderton C, Pranata A (2024) Digital interventions for healthy ageing and cognitive health in older adults: a systematic review of mixed method studies and meta-analysis. BMC Geriatr 24(1):217. https://doi.org/10.1186/s12877-023-04617-3
46. Huynh Q et al (2025) Barriers to and facilitators of digital health technology adoption among older adults with chronic diseases: updated systematic review. JMIR Aging 8:80000. https://doi.org/10.2196/80000
47. Lestari HM, Miranda AV, Fuady A (2024) Barriers to telemedicine adoption among rural communities in developing countries: a systematic review and proposed framework. Clinical Epidemiology and Global Health 28:101684. https://doi.org/10.1016/j.cegh.2024.101684
48. Şahin E, Yavuz Veizi BG, Naharci MI (2024) Telemedicine interventions for older adults: a systematic review. J Telemed Telecare 30(2):305–319. https://doi.org/10.1177/1357633x211058340
49. Mathew S et al (2025) Telehealth for primary healthcare delivery in rural and remote contexts in high-income countries: a scoping review. mHealth. https://doi.org/10.21037/mhealth-24-42. Published online 30 June 2025
50. World Health Organization (2021) Global Strategy on Digital Health 2020–2025. https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf. Accessed 01 May 2025
51. Health TLD (2023) Digital health equity for older populations. The Lancet Digital Health 5(7):395. https://doi.org/10.1016/S2589-7500(23)00114-0
52. Hatef E, Weiner JP et al (2025) Digital health care equity framework (DHEF). JAMIA Open. https://doi.org/10.1093/jamiaopen/ooae153
53. O'Callaghan ME et al (2023) Healthcare inequity and digital health - a bridge for the divide, or further erosion of the chasm? PLOS Digital Health 2(6):0000268. https://doi.org/10.1371/journal.pdig.0000268
54. Wani S et al (2024) Security and privacy challenges, issues, and enhancing techniques for internet of medical things: a systematic review. Security and Privacy 7(5):409. https://doi.org/10.1002/spy2.409
55. Al-Saleem M et al (2026) Security and privacy challenges in healthcare IoT systems with cloud, edge, and fog computing: a systematic literature review. Comput Sci Rev. https://doi.org/10.1016/j.cosrev.2026.100739. Published online June 2026
56. Abubakar M et al (2026) A comprehensive survey on security and privacy in IoMT-driven smart healthcare. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1007/s44443-026-00586-w
57. Gupta R et al (2026) A systematic review of emerging trends of IoT in healthcare and IoMT frameworks. Discover Public Health. https://doi.org/10.1186/s12982-026-01347-w
58. Triantaphyllou E, Sánchez A (1997) Determining the most critical criteria in deterministic and fuzzy decision making. Comput Ind Eng 32(2):299–311. https://doi.org/10.1016/S0360-8352(96)00213-9
59. García-Zamora D, Dutta B, Figueira JR, Martínez L (2024) The deck of cards method to build interpretable fuzzy sets in decision-making. Eur J Oper Res 318(3):816–834. https://doi.org/10.1016/j.ejor.2024.07.003
60. Dutta B, García-Zamora D, Figueira JR, Martínez L (2026) A deck of cards-based co-constructive approach for modeling higher-order uncertainty in fuzzy decision-making. Appl Soft Comput 196:115094. https://doi.org/10.1016/j.asoc.2026.115094

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.