Measuring Interdisciplinarity: A Multi-Component Indicator Panel for Research Evaluation

A. Rivero

2026

Abstract

We review the landscape of bibliometric indicators for measuring interdisciplinary research and propose a structured three-component panel that combines Rao-Stirling diversity, network coherence (mean linkage strength), and a cross-field effect proxy. Drawing on recent work that exposes the low consistency and limited construct validity of existing single-indicator approaches, we organize the literature into a taxonomy of four conceptual dimensions — diversity, coherence, diffusion, and novelty — and four methodological families. Using a toy university dataset, we demonstrate that no single scalar indicator can discriminate genuine cross-disciplinary integration from polymathic breadth or narrow specialization. The full panel, by contrast, uniquely characterizes each researcher type. We prove analytically that this discrimination is robust under perturbation of the inter-category similarity matrix. A department-level case study with seven researchers confirms the panel's discriminatory power at realistic scale, correctly distinguishing integrators, polymaths, and bridge specialists where single indicators fail. These results suggest that evaluation of interdisciplinary research — whether at the institutional or national agency level — requires a multidimensional approach rather than composite scoring.

Introduction

Interdisciplinary research (IDR) is widely regarded as essential for addressing complex scientific and societal challenges. Policy initiatives in the United States, Europe, and elsewhere have sought to foster interdisciplinarity, often predicated on the assumption that crossing disciplinary boundaries leads to more impactful research outcomes (National Academies, 2005). However, the measurement of interdisciplinarity remains problematic. Despite decades of work, the bibliometric literature has not converged on appropriate indicators.

Wang and Schneider (2020), testing 23 interdisciplinarity measures across four methodological families, found surprisingly low correlations even among measures designed to capture the same dimension, concluding that "no single indicator can unequivocally identify" interdisciplinary research. Leydesdorff, Wagner,

and Bornmann (2019) showed that the widely used Rao-Stirling diversity index is dominated by its disparity component, producing anomalous rankings and low discriminatory power. Cantone (2024) argued that interdisciplinarity is a "polysemous construct" whose multiple semantic dimensions cannot be captured by any single numerical value. These findings point to a fundamental mismatch between the multidimensional nature of IDR and the scalar indicators used to measure it.

This paper makes two contributions. First, we survey the indicator landscape and organize it into a taxonomy of four conceptual dimensions (diversity, coherence, diffusion, novelty) crossed with four methodological families (reference-based, citation-based, text-based, network-based). Second, we propose a specific three-component panel — diversity, coherence, and cross-field effect — and demonstrate its discrimination power on a toy dataset with analytical robustness guarantees.

The remainder of the paper is structured as follows. Section 2 reviews the conceptual foundations of interdisciplinarity measurement. Section 3 presents the indicator taxonomy. Section 4 defines our panel, demonstrates its discrimination power, and proves robustness results. Section 5 discusses implications for institutional and national-level evaluation. Section 6 identifies open problems. Section 7 applies the panel to a department-level case study. Section 8 concludes.

Conceptual Foundations

Stirling's Diversity Framework

The difficulty of measuring interdisciplinarity is fundamentally a diversity measurement problem. Stirling (2007) demonstrated that any meaningful characterization of diversity requires attention to three properties — *variety* (how many categories are represented), *balance* (how evenly distributed they are), and *disparity* (how different the categories are from each other) — and that standard indices such as Shannon entropy or the Herfindahl concentration index capture only the first two. This tripartite framework has become the standard conceptual lens through which interdisciplinarity indicators are analyzed in the bibliometric literature (Porter and Rafols, 2009; Leydesdorff and Rafols, 2011; Wang and Schneider, 2020). The framework's influence extends beyond bibliometrics: Stirling originally developed it in ecological and technological diversity contexts, and its adoption by the scientometrics community reflects a recognition that interdisciplinarity, like biodiversity, cannot be reduced to a count of categories without attending to the distances between them. The Rao-Stirling index $\Delta = \sum_{i \neq j} d_{ij} p_i p_j$, which operationalizes all three properties in a single expression, has consequently become the most widely used point of departure for indicator design (Rafols and Meyer, 2009; Zhang, Rousseau, and Glanzel, 2016).

Historical Evolution

The modern study of interdisciplinarity has roots in science policy debates of the 1970s, but quantitative measurement efforts accelerated only after the OECD codified a tripartite typology that remains influential today (OECD, 1998; Morillo, Bordons, and Gomez, 2003). Under that typology, *multidisciplinary* research draws on different disciplinary perspectives without integrating them; *interdisciplinary* research achieves a coherent theoretical, conceptual, or methodological synthesis; and *transdisciplinary* research entails a mutual integration of disciplinary epistemologies that may transcend existing boundaries altogether. While the boundaries between these categories remain contested, the distinction foregrounds a critical question for indicator design: should measurement target the breadth of disciplinary inputs (a multidisciplinary property) or the depth of their integration (an interdisciplinary or transdisciplinary property)?

Wagner et al. (2011) elaborate these definitions in a comprehensive review of IDR measurement. Multidisciplinary research "juxtaposes disciplinary perspectives, adding breadth and available knowledge — the product is no more and no less than the simple sum of its parts." Interdisciplinary research "integrates separate disciplinary data, methods, tools, concepts, and theories in order to create a holistic view" — the product is "different from, and greater than, the sum of its parts." Transdisciplinary approaches "are comprehensive frameworks that transcend the narrow scope of disciplinary worldviews." They note that common usage "rarely distinguishes between the input and output directions of IDR" (Wagner et al., 2011, fn. 10), yet the distinction matters for measurement: a researcher may draw on multiple disciplines' methods in designing a study (input-side interdisciplinarity) while publishing exclusively in one field (output-side concentration), or conversely may publish across many fields without methodological integration. A fourth mode, *cross-disciplinary* research, involves referencing literature from another field without any attempt at integration (Aksnes, Karlstrøm, and Piro, 2026; Hammarfelt, 2020).

Aksnes, Karlstrøm, and Piro (2026), surveying 1,498 publications with self-reported IDR ratings, found that 42 percent of papers were rated as *both* multidisciplinary and interdisciplinary, and a further 23 percent as partially both. This empirical overlap confirms that the multi/inter distinction is not a clean partition but a spectrum, and that single-scalar diversity indicators (which aggregate breadth without regard for integration) cannot distinguish between these modes. As Choi and Pak (2006) put it via Abramo, D'Angelo, and Zhang (2018), multi-, inter-, and transdisciplinarity form "a continuum of increasing levels of involvement by multiple disciplines." The directionality of knowledge flow matters as well. When the output of one discipline serves as input for another without synthesis, the literature terms this *sequential multidisciplinarity* (Stokols et al., 2003); when the borrowed input transforms the receiving discipline, it is *instrumental interdisciplinarity* (Klein, 2008). Bidirectional exchange constitutes *reciprocal interdisciplinarity*. Zhou, Guns, and Engels (2023) formalize the flow perspective through their interdisciplinary knowledge flow (IKF) framework,

which characterizes exchanges along three dimensions: broadness, intensity, and homogeneity. These distinctions have direct implications for the panel introduced in Section 4: diversity (Δ) captures breadth regardless of direction, coherence (S) distinguishes integration from juxtaposition, and the cross-field effect (E) measures diffusion beyond the home discipline.

Empirical evidence on long-term trends sharpens this question. Porter and Rafols (2009) analyzed publication records across ten subject areas from 1975 to 2005 and documented pervasive growth in surface-level markers of interdisciplinarity: the average number of authors per paper increased by roughly 75 percent (from 1.3 to 2.0 in mathematics, 3.0 to 6.1 in medical research and education), the average number of references per paper grew by approximately 50 percent, and the diversity of cited disciplines expanded comparably. Single-author publication rates declined sharply across all fields (from 71 to 37 percent in mathematics; from 40 to 20 percent in physics and biology; from 12 to 4 percent in chemistry). Yet integration scores based on the Rao-Stirling index showed only a modest average increase of roughly 5 percent over the same period (excluding mathematics, where the increase reached 39 percent from a very low base). The conclusion was striking: science was "becoming more interdisciplinary, but in small steps," with citations mainly reaching neighboring fields and only modest growth in distant cognitive connections.

Morillo, Bordons, and Gomez (2001, 2003) provided complementary evidence at the journal and category level. Between 1981 and 1996, the ISI journal classification system added 38 new subject categories, of which 21 appeared in engineering alone — a 154 percent increase in that field's journal count. The new categories exhibited systematically higher interdisciplinarity: 69 percent were multi-assigned to more than one category (compared to 55 percent of older categories), and they showed 28 percent stronger inter-category link strength and 30 percent greater disciplinary diversity. Fully 80 percent of the new categories fell into clusters characterized by high interdisciplinarity. This pattern suggests that the growth of science proceeds through simultaneous fragmentation and hybridization — new specialties emerge at disciplinary boundaries and inherit an interdisciplinary character from their origins.

The Polysemy Problem

Interdisciplinarity itself is not a unitary concept. Cantone (2024) emphasizes that IDR is a "polysemous construct" — a term that carries multiple, partially overlapping meanings across scholarly communities and policy contexts. A physicist collaborating with biologists, a data scientist applying methods across domains, and a social scientist synthesizing theories from multiple disciplines are all called "interdisciplinary," yet the nature and depth of their boundary-crossing differ qualitatively. This polysemy means that any measurement system must either select a specific operational definition of interdisciplinarity or explicitly accommodate multiple dimensions. The taxonomy presented in Section 3 follows the latter strategy.

The polysemy problem is compounded by a cognitive–social distinction that cuts across all definitions (Cantone, 2024). Cognitive approaches measure interdisciplinarity through the diversity of knowledge inputs — references, methods, theoretical frameworks — and thus track epistemological breadth. Grouping approaches measure it through social structures — co-authorships, institutional affiliations, panel compositions. Semantic approaches analyze textual content — keywords, abstracts, full text — to detect topical boundary-crossing. Each captures a different facet of the phenomenon, and high scores on one dimension need not correlate with high scores on another. The practical consequence is that researchers who self-identify as interdisciplinary may not register as such on bibliometric indicators, and vice versa (Zwanenburg, 2022). This discordance between self-reported and measured interdisciplinarity is not merely a calibration problem; it reflects genuinely different construals of what the term means.

Morillo, Bordons, and Gomez (2003) introduced a further distinction between "big interdisciplinarity" — connections between distant disciplinary areas — and "small interdisciplinarity" — connections between neighboring categories within the same broad field. Applied Chemistry, for instance, exhibited 83 percent multi-assigned journals and 55.6 percent external (cross-area) links, while Polymer Science showed only 39 percent multi-assignation and 33.3 percent external links. The big/small distinction has direct implications for indicator design: an index sensitive only to variety will conflate these two qualitatively different patterns, whereas one that incorporates disparity will distinguish them.

Epistemological Challenges

Beyond polysemy, the measurement of interdisciplinarity confronts several epistemological difficulties that constrain what indicators can legitimately claim to capture. The most fundamental is the instability of the disciplinary classification systems on which all bibliometric indicators depend. The ISI Web of Science subject categories — the most widely used classification — achieve only approximately 50 percent alignment with citation-based cluster solutions (Boyack, 2005), and the match with network-derived classifications is similarly imperfect (Leydesdorff, 2006). As of recent counts, the system comprises 254 subject categories, with 39 percent of journals assigned to more than one category. While Rafols and Leydesdorff showed that these misalignments have limited effects on aggregate science maps, they can substantially affect individual-level indicator values.

A related problem concerns "horizontal" disciplines — broad categories such as Biology, Physics, Chemistry, or the catch-all Multidisciplinary Sciences — which exhibit artificially low multi-assignation precisely because journal classification policies limit excessive dispersion across categories (Morillo, Bordons, and Gomez, 2003). Journals like Nature, Science, and PNAS appear in Multidisciplinary Sciences yet produce low interdisciplinarity scores under standard indicators despite publishing work that spans the entire disciplinary spectrum. This signals a deeper epistemological concern: the categories we use to define disciplinary

boundaries are themselves artifacts of administrative and historical convention, not stable features of the knowledge landscape (Cantone, 2024). Any indicator built on such classifications inherits their contingency.

Conceptual vs. Empirical Validity

A further complication is the gap between conceptual definitions and empirical operationalization. An indicator may have a clear theoretical motivation (e.g., "diversity of knowledge inputs") yet fail to discriminate meaningfully when applied to real data. Wang and Schneider (2020) documented this problem systematically: measures that should be theoretically equivalent produce inconsistent and sometimes contradictory rankings when applied to the same dataset. Leydesdorff et al. (2019) showed that Rao-Stirling diversity values often differ only at the third decimal place across researchers, limiting practical discriminatory power. These findings underscore the need for empirical validation of any proposed indicator, not merely theoretical justification.

The validity crisis has multiple dimensions (Zwanenburg, 2022). *Content validity* is threatened by the disparity between the conceptual richness of interdisciplinarity — which encompasses variety, balance, and disparity — and the tendency of individual indicators to capture only one or two of these facets. *Domain validity* is undermined by the multiplicity of classification systems and the ambiguity of journal-to-category allocations. *Coherence validity* requires consistency across alternative operationalizations of the same construct, yet empirical studies repeatedly find that notionally equivalent measures produce divergent results. Rafols (2019) argued that most existing indicators lack either analytical validity (they do not measure what they claim to measure) or social robustness (they are not perceived as meaningful by the communities they evaluate), and that responsible metrics require both. The principles of robustness, humility, transparency, and sensitivity to epistemic diversity that Rafols articulated provide a normative framework that any proposed measurement system should satisfy.

A distinct but related problem is the confounding of interdisciplinarity with research quality. Indicators do not record a pre-existing property of "quality" or "excellence"; rather, as Rafols (2019) argued, they *enact* these categories — outside assessment practices, such properties have no independent existence. When interdisciplinarity indicators are used alongside or in combination with citation-based quality proxies, the risk of circular reasoning is acute: interdisciplinary work may receive higher citations precisely because it reaches broader audiences, not because it is intrinsically superior. The panel proposed in Section 4 treats quality as orthogonal to interdisciplinarity characterization, following the principle that the panel should describe the *type* of boundary-crossing without adjudicating its merit.

Methodological Pluralism

The diversity of conceptual perspectives reviewed above has given rise to a correspondingly diverse toolkit of measurement approaches. At the most established end, the Rao-Stirling diversity index captures variety, balance, and disparity in a single expression, distinguishing it from Shannon entropy and the Herfindahl index, which incorporate no measure of inter-category distance (Porter and Rafols, 2009). At the journal and category level, the Salton cosine index normalizes shared journal counts between two categories by the geometric mean of their sizes, providing a symmetric measure of inter-category link strength that ranges from zero to one (Morillo, Bordons, and Gomez, 2003). Multi-assignation patterns — whether a journal's secondary categories fall within the same broad area (internal, or "small" interdisciplinarity) or across areas (external, or "big" interdisciplinarity) — offer a complementary structural perspective.

Different indicators are appropriate at different levels of aggregation. Journal multi-assignation in the ISI system provides a macro-level view that is easy to apply but coarse in resolution. JCR citation and reference patterns operate at the category level and are more sensitive to disciplinary dynamics. Detailed section-level analyses (e.g., using Chemical Abstracts sections) offer journal-level precision but require domain-specific infrastructure (Morillo, Bordons, and Gomez, 2001). More recent proposals advocate semi-qualitative, contextual methods — including overlay maps on a base science map that visualize variety, balance, and disparity simultaneously — as alternatives to or complements for scalar indices (Rafols, 2019). The guiding principle is that of "indicators in the plural": no single metric suffices, and responsible evaluation requires triangulation across methods and levels of analysis. The multi-component panel developed in Section 4 is designed in this spirit.

A Taxonomy of Interdisciplinarity Indicators

We organize the indicator literature along two axes: the *conceptual dimension* of interdisciplinarity being measured, and the *methodological family* of the indicator. This yields a structured map of the field that clarifies what each indicator actually captures and where gaps remain.

Diversity Indicators

Diversity indicators measure the heterogeneity of a researcher's knowledge inputs — typically the disciplinary spread of cited references. They are the most extensively studied family of interdisciplinarity measures, and any credible assessment of IDR at the paper, author, or institutional level must reckon with the conceptual and computational choices embedded in their design. This subsection provides a systematic treatment of the principal index families, the similarity matrices they require, the empirical evidence on their validity, and the practical difficulties that arise when one applies them to real bibliometric data.

Stirling's three-property framework

The theoretical foundation for most modern diversity indicators is Stirling's (2007) decomposition of diversity into three properties: *variety* (the number of distinct categories to which elements are assigned), *balance* (the evenness of the distribution of elements across those categories), and *disparity* (the degree of difference between the categories themselves). A fully satisfactory diversity measure should be sensitive to all three properties simultaneously. In practice, however, many widely used indices capture only one or two of the three, which is a principal source of disagreement among empirical studies.

Measures capturing variety and balance only

Several classical indices from ecology and economics have been adapted for IDR measurement. These indices register variety and balance but are blind to disparity — they treat a shift from Organic Chemistry to Analytical Chemistry identically to a shift from Organic Chemistry to Sociology.

Shannon entropy. For a publication whose references fall in categories $i = 1, \ldots, n$ with proportions p_i, the Shannon entropy is $H = -\sum_i p_i \ln p_i$. It reaches its maximum $\ln n$ when references are spread uniformly across n categories and equals zero when all references fall in a single category.

Simpson diversity. The Simpson index $D_{\text{Sim}} = 1 - \sum_i p_i^2$ gives the probability that two references drawn at random belong to different categories. It is bounded between 0 and $1 - 1/n$ and is more sensitive to dominant categories than Shannon entropy.

Brillouin index. A finite-sample analogue of Shannon entropy, defined as $HB = [\log(\sum_i c_i)! - \sum_i \log c_i!] / \sum_i c_i$ where c_i is the count of references in category i. Wang and Schneider (2020) found Shannon and Brillouin to be nearly perfectly correlated ($r = 1.00$ in their Table 5), rendering them empirically redundant.

Inverted Gini coefficient. The Gini coefficient measures concentration; $1 - G$ converts it into a balance indicator. Like Shannon, it is insensitive to the identity of the categories across which references are distributed.

An important empirical regularity is that these non-disparity indices are moderately to strongly correlated among themselves (Wang and Schneider, 2020, Table 5, with pairwise Pearson correlations between Simpson, Shannon, Brillouin, and $1 -$ Gini ranging from 0.60 to 1.00), but only weakly correlated with disparity-incorporating indices. The two families thus capture genuinely different aspects of diversity.

The Rao-Stirling diversity index

The Rao-Stirling index is the most widely used indicator that incorporates all three of Stirling's properties (Porter and Rafols, 2009; Rafols and Meyer, 2009).

In its standard form it is defined as

$$\Delta = \sum_{\substack{i,j \\ i \neq j}} d_{ij}\, p_i\, p_j$$

where p_i is the proportion of references in category i and $d_{ij} = 1 - s_{ij}$ is the dissimilarity between categories i and j, derived from a similarity matrix $\mathbf{S} = [s_{ij}]$. When all categories are maximally dissimilar ($d_{ij} = 1$ for all $i \neq j$), the Rao-Stirling index reduces to the Simpson index. When the similarity matrix carries real structure, the index penalises diversity among close categories and rewards diversity among distant ones.

Alpha-beta generalisation. Stirling (2007) proposed a more general family $D_{\alpha,\beta} = \sum_{i \neq j} d_{ij}^{\alpha}\, (p_i\, p_j)^{\beta}$, where the exponents α and β govern the relative weight given to disparity versus balance. The conventional Rao-Stirling choice sets $\alpha = \beta = 1$. Increasing α amplifies the contribution of highly disparate category pairs; increasing β amplifies the contribution of well-balanced distributions. In most empirical scientometric work $\alpha = \beta = 1$ is adopted without discussion, but the sensitivity of results to these parameter choices has received limited investigation. Researchers should be aware that changing these exponents can shift rankings of papers or fields, even when the underlying data are identical.

The eight-variant study. Wang and Schneider (2020) tested eight variants of the Rao-Stirling index by crossing two classification-level choices (individual-publication average RS_P versus aggregated RS_G) with four dissimilarity-matrix specifications (using either the Salton vector cosine SC or the Ochiai scalar cosine SO, each converted to dissimilarity by either $1 - s$ or $1/s$). The results are sobering. Pearson correlations between variants using the same cosine formula but different dissimilarity transformations can be as low as $r = 0.30$ (between RS_G$[1 - S_C]$ and RS_G$[1/S_C]$); correlations between variants using different cosine formulas with the same transformation can be as low as $r = 0.18$ (between RS_P$[1 - S_C]$ and RS_P$[1 - S_O]$). In an in-depth analysis of five selected Web of Science subject categories (Nanoscience, Biochemistry, Library and Information Science, Law, and Mathematics), the rankings produced by different variants frequently contradicted one another. For instance, Mathematics was ranked 221st out of 224 categories by RS_P$[1 - S_C]$ but 79th by RS_G$[1 - S_C]$, despite the strong overall Spearman correlation ($\rho = 0.91$) between these two variants. Wang and Schneider concluded that "the current measurements of interdisciplinarity should be interpreted with much caution."

Hill-type true diversity measures

Zhang, Rousseau, and Glanzel (2016) proposed Hill-type measures adapted from the ecological diversity literature (Hill, 1973; Jost, 2006; Leinster and Cobbold, 2012). The general form is

$$ {}^{q}D^{S} = \left(\sum_{i=1}^{N} p_i \left(\sum_{j=1}^{N} s_{ij}\, p_j \right)^{q-1} \right)^{1/(1-q)} $$

where $\mathbf{S} = [s_{ij}]$ is a similarity matrix with $s_{ii} = 1$ and $0 \leq s_{ij} = s_{ji} \leq 1$, and q is a sensitivity parameter. The special case $q = 2$ yields

$$ {}^{2}D^{S} = \frac{1}{\sum_{i,j} s_{ij}\, p_i\, p_j} $$

which is related to, but distinct from, the Rao-Stirling index: substituting $d_{ij} = 1 - s_{ij}$, one obtains ${}^{2}D^{S} = 1/(1 - \Delta)$. While the Rao-Stirling index is bounded between 0 and 1, ${}^{2}D^{S}$ ranges from 1 (a single category, or perfectly similar categories) to N (all categories equally abundant and maximally dissimilar), which confers on it the interpretation of an "effective number of disciplines."

The mathematical advantage of Hill-type measures is that they satisfy six desirable properties that entropy-based indices violate (Jost, 2006, 2009): symmetry, zero-output independence, the transfer principle, scale invariance, the replication principle, and normalisation. The replication principle is especially important for policy discussions: if m equally diverse, non-overlapping research portfolios are pooled, a true diversity measure should give the pooled portfolio a diversity of $m \cdot D_0$. Shannon entropy and the Simpson index fail this test. Only when working with true diversities does it make sense to discuss percentage changes in diversity — a property that makes Hill-type measures particularly attractive for longitudinal and comparative studies. In an empirical demonstration, Zhang, Rousseau, and Glanzel (2016) showed that ${}^{2}D^{S}$ discriminates more effectively than the Rao-Stirling index among journals spanning a range from specialised mathematics to multidisciplinary science.

Multi-assignation and the Morillo-Bordons-Gomez approach

An entirely different tradition constructs interdisciplinarity indicators from the multi-assignation of journals to classification categories rather than from reference-list analysis. Morillo, Bordons, and Gomez (2001) introduced a suite of indicators for the ISI (now Clarivate) subject-category system: the percentage of multi-assigned journals in a category, the percentage of journals assigned to categories outside the research area, and the concentration of references across categories (measured via the Pratt index). These indicators were validated through a case study in Chemistry, where Applied Chemistry — a "horizontal" discipline with 83% multi-assigned journals — consistently scored higher than Polymer Science (39% multi-assigned) across all indicators.

In a follow-up study, Morillo, Bordons, and Gomez (2003) extended the analysis to all ISI categories and established a four-cluster typology of disciplines based

on multi-assignation percentage, external link percentage, diversity of links, and strength of links (the last measured by the Salton cosine of shared journal sets). Two of the four clusters were characterised as reflecting, respectively, "big interdisciplinarity" — in which links connect categories from different research areas (e.g., Biotechnology linking Life Sciences and Engineering) — and "small interdisciplinarity" — in which links remain within the same area (e.g., Transplantation linking Surgery and Immunology). This distinction between distant-category and close-category integration anticipated the later emphasis on disparity in the Stirling framework.

These multi-assignation indicators have distinct advantages: they are straightforward to compute, do not require a similarity matrix, and can be applied at the macro level (research areas) as well as the meso level (categories and journals). Their disadvantage is sensitivity to the ISI classification scheme itself — particularly for "horizontal" categories such as Chemistry or Physics, where ISI artificially limits multi-assignation. They also do not operate at the individual-paper level, which limits their applicability in researcher-level evaluation.

Similarity matrix estimation

All disparity-sensitive measures depend on a matrix $\mathbf{S}$ (or its complement $\mathbf{D} = \mathbf{1} - \mathbf{S}$) encoding pairwise relationships among classification categories. The standard approach constructs $\mathbf{S}$ from inter-category citation flows using a cosine similarity. Historically, this construction extends journal-journal citation mapping methods developed for JCR-scale classification work (Leydesdorff, 2006). Wang and Schneider (2020) distinguished two variants: the Salton vector cosine $SC(i,j) = \sum_k c_{ik}\, c_{jk} / \sqrt{\sum_k c_{ik}^2 \cdot \sum_k c_{jk}^2}$, which compares the citing profiles of two categories, and the Ochiai scalar cosine $SO(i,j) = (c_{ij} + c_{ji}) / \sqrt{(\sum_k c_{ik} + \sum_k c_{ki})(\sum_k c_{jk} + \sum_k c_{kj})}$, which uses direct bilateral citation exchange. These two cosine formulations can yield substantially different similarity landscapes: the SO-based matrices are extremely left-skewed (most pairs have dissimilarity close to 1), which means that SO-based Rao-Stirling variants approach the Simpson index in practice.

A further degree of freedom is the transformation from similarity to dissimilarity. The standard choice $d_{ij} = 1 - s_{ij}$ is intuitive but not the only option; $d_{ij} = 1/s_{ij}$ has also been used (Jensen and Lutkouskaya, 2014). Wang and Schneider (2020) showed that the combination of cosine variant and dissimilarity transformation accounts for more variance in Rao-Stirling outcomes than any other single methodological decision.

The requirement for a large citation database is itself a barrier to entry. Cantone, Zheng, Tomaselli, and Nightingale (2025) recently proposed an alternative: estimating similarity matrices directly from large language models (LLMs). In their experimental protocol, ChatGPT, Gemini, and Claude were each prompted to produce numerical similarity estimates for pairs of disciplines under two

taxonomies. Across 228 sampled matrices (16,200 individual estimates), they evaluated five properties: precision (inverse variance across repeated identical prompts), agreement (cross-model correlation), resilience (sensitivity to semantically trivial rewording of discipline names), robustness (sensitivity to reordering), and explainability. Gemini achieved estimates closest to traditional citation-based matrices; Claude showed a balanced profile; and ChatGPT exhibited superior resilience to prompt variation. While none of the models reached perfect agreement with citation-based baselines, the authors concluded that LLM-based estimation is "sufficiently well" suited for the task and offers a low-cost, database-free alternative that could democratise access to disparity-sensitive IDR measurement.

Normalization, calibration, and classification granularity

A pervasive difficulty with all diversity indicators is their dependence on the classification system. A publication may appear more interdisciplinary under the 250-category Web of Science scheme than under the 40-category OECD scheme simply because finer granularity creates more category boundaries to cross. Zhang, Rousseau, and Glanzel (2016) demonstrated this directly: journal rankings by ${}^2D^S$ shifted substantially when measured against the 68 ECOOM subfields versus the 16 ECOOM major fields, with Spearman correlations between the two rankings of only 0.79. This classification sensitivity is dampened, though not eliminated, when the measure incorporates disparity — splitting a field into subfields produces subfields that are similar, and their contribution to disparity-sensitive indices is accordingly attenuated (Zwanenburg, Nakhoda, and Whigham, 2022).

Field normalization raises further complications. A paper in Mathematics citing three categories may represent greater knowledge breadth, relative to disciplinary norms, than a paper in Biomedicine citing ten. Without baseline correction, raw diversity scores penalise fields that are naturally specialised and reward fields that are inherently diffuse. Leydesdorff, Wagner, and Bornmann (2019) proposed a decomposed diversity measure (DIV) as the product of normalised variety, balance ($1 - \text{Gini}$), and disparity components, allowing each component to be inspected separately. Whether these components should be combined multiplicatively (Leydesdorff, Wagner, and Bornmann, 2019) or additively (Mutz, 2022) remains an open question that the definition of IDR alone does not resolve (Zwanenburg, Nakhoda, and Whigham, 2022).

Validity assessment: the Zwanenburg evaluation

The most systematic validity assessment to date is that of Zwanenburg, Nakhoda, and Whigham (2022), who evaluated 21 measures of IDR against eight criteria derived from a synthesis of 25 definitions of interdisciplinarity. The eight criteria are organised under four headings:

1. *Applicability*: (1a) Multi-object — the measure should be applicable to pa-

pers, authors, institutions, journals, and disciplines; (1b) Size independence — scores should not vary merely because the object of study represents more or fewer publications.

2. *Integration evidence*: (2) The measure should be based on evidence that knowledge is actually integrated, not merely that multiple disciplines are represented. Paper-level reference analysis provides adequate evidence; aggregated journal-level citation counts do not.
3. *Discipline identification*: (3a) Valid allocation of references to disciplines; (3b) Identification of all source disciplines (completeness); (3c) Low classification bias — the measure should not produce wildly different scores when applied to classifications of different granularity.
4. *Diversity capture*: (4a) Sensitivity to all three diversity dimensions (variety, balance, disparity); (4b) Decomposability into separate scores for each dimension.

Of the 21 measures evaluated, only six met the criterion for evidence of knowledge integration (criterion 2): the Rao-Stirling index, the Hill-type measure, the Coherence measure, the DIV indicator, the overall diversity indicator (d_{ive}), and the Reverse Simpson Index applied at the paper level. The remaining 15 either operated at aggregate levels that preclude integration evidence, or relied on classification overlap rather than citation-based evidence. Within the six that met criterion 2, the Rao-Stirling and Hill-type measures also satisfied criterion 4a (all diversity aspects captured) and criterion 3c (low classification bias), but neither was decomposable into separate variety, balance, and disparity scores (criterion 4b). Only the DIV measure and the overall diversity indicator d_{ive} met all four of criteria 1a, 1b, 2, 3c, 4a, and 4b.

Zwanenburg, Nakhoda, and Whigham cautioned that no single measure satisfied all eight criteria, and that the criterion for valid discipline allocation (3a) remained unresolved for every measure relying on journal-to-category mappings — approximately 30% of references in their institutional database were assigned to multiple WoS categories, creating allocation ambiguities that cascade into inflated diversity scores.

Author versus reference diversity

The choice of *what* to diversify introduces a further dimension. Most indicators measure diversity over the reference list, but an alternative tradition measures diversity over the disciplinary affiliations of co-authors. Abramo, D'Angelo, and Zhang (2018) compared the two approaches using 43,667 Italian university publications, partitioned into single-author papers (by construction non-interdisciplinary under the author method), multi-author single-field papers, and multi-field papers. They found general convergence: reference-list diversity increased with the number of distinct disciplinary sectors (SDSs) reflected in the byline, and disparity was higher when the authors' SDSs spanned different university disciplinary areas (UDAs) rather than the same UDA. However, striking individual exceptions emerged. The three publications with the highest inte-

grated diversity score in the entire dataset were single-author papers, precisely the "intrapersonal integrators" whose knowledge breadth cannot be detected by the author method. This finding highlights a structural limitation of author-based approaches and underscores the complementary value of reference-based analysis for identifying individual knowledge integration.

Summary

The landscape of diversity indicators is rich but fragmented. Measures that incorporate pairwise dissimilarity (Rao-Stirling, Hill-type, DIV) form one empirical cluster; measures that do not (Shannon, Simpson, multi-assignation counts) form another, and inter-group correlations are weak (Wang and Schneider, 2020). The choice of similarity matrix specification accounts for as much variance in outcomes as the choice of index formula. No single measure satisfies all validity criteria (Zwanenburg, Nakhoda, and Whigham, 2022), and LLM-based similarity estimation opens a promising but still immature alternative to citation-derived matrices (Cantone, Zheng, Tomaselli, and Nightingale, 2025). For applied evaluation, the implication is clear: diversity indicators should be reported alongside their computational specifications (classification scheme, similarity method, dissimilarity transformation, aggregation level), and conclusions that depend on a single indicator variant should be treated with caution.

Coherence Indicators

Diversity indicators, however richly specified, answer only one question: *how heterogeneous are the knowledge inputs?* They are silent on whether those heterogeneous inputs have been woven into a unified intellectual fabric or merely placed side by side. Rafols and Meyer (2009) introduced the concept of *coherence* to fill this gap, defining it as "the extent to which specific topics, concepts, tools, data, etc. used in a research process are related" (p. 175). Whereas diversity captures the categorical breadth of references, coherence captures the relational structure among the items within those categories — the intensity of their mutual integration.

The distinction matters for evaluation. High diversity alone does not guarantee that disparate knowledge sources have been synthesized; it may reflect mere juxtaposition or polymathic breadth across unrelated literatures. Coherence supplies the missing signal. Moreover, the functional interpretation of coherence depends on the unit of analysis. High coherence in an article's reference list indicates that the article builds on an established, internally connected specialty. High coherence across a research centre's publications indicates that the centre is achieving its integrative mission. Low coherence, conversely, signals that previously unrelated bodies of knowledge are being brought into contact — a state of *potential* interdisciplinary integration that may mature over time.

Bibliographic coupling operationalization

Rafols and Meyer (2009) operationalized coherence through bibliographic coupling: two publications are linked to the extent that they share references, and the density of the resulting network serves as a coherence indicator for the set. The similarity between any two publications a and b is computed using Salton's cosine,

$$s_{ab} = \frac{\mathbf{r}_a \cdot \mathbf{r}_b}{\|\mathbf{r}_a\| \, \|\mathbf{r}_b\|}$$

where $\mathbf{r}_a$ is the binary reference vector of publication a. Cosine normalization controls for the total number of references in each publication, a desirable property that guards against size-driven artefacts.

From the resulting similarity matrix two network-level coherence indicators are derived:

Mean linkage strength (S). The mean of the off-diagonal entries of the normalized bibliographic coupling matrix,

$$S = \frac{2}{N(N-1)} \sum_{a<b} s_{ab}$$

where N is the number of publications. In a binary network S reduces to ordinary network density; in a valued network it captures both the proportion of realized links and their average intensity. S is bounded between 0 and 1 and was found to be scale-invariant across network sizes ranging from 10 to 1,275 nodes in Rafols and Meyer's kinesin benchmark sample.

Mean path length (L). The average shortest-path distance between all pairs of nodes in the binarized similarity network. Binarization requires a threshold τ below which pairwise similarities are treated as zero; Rafols and Meyer adopted $\tau = 0.05$ (equivalent to requiring at least one shared reference in a 20-reference bibliography) to suppress spurious links arising from highly cited general references. Lower values of L indicate a more compact, internally connected body of work. In their molecular-motors sample, S and L were highly correlated ($r \approx 0.95$), suggesting that the two indicators capture essentially the same structural property and that S alone may suffice in many applications.

The choice of bibliographic coupling — rather than co-citation — as the underlying relation is deliberate. Bibliographic coupling is forward-looking: it reflects the knowledge sources that authors chose to draw upon at the time of writing, rather than the audience patterns that emerge after publication. This makes it applicable to recent publications for which a citation window has not yet accumulated.

Empirical evidence: orthogonality from diversity

Rafols and Meyer (2009) tested their coherence indicators on 12 articles drawn from the molecular motors literature. Diversity and coherence were found to be uncorrelated — the two dimensions offered "orthogonal perspectives" on interdisciplinarity. Coherence values spanned a wide range (S from 0.024 to 0.113). At one extreme, Noji (1997) exhibited $S = 0.024$: its reference network showed a clear divide between the bioenergetics and linear-motor literatures, connected only through a single review article. This low coherence signalled a seminal act of integration in which two previously separate research strands were being brought into contact for the first time. At the other extreme, Tomishige (2002) exhibited $S = 0.113$: it drew on what had by then become an established interdisciplinary specialty, and its references formed a dense, internally connected cluster.

Wang and Schneider (2020) confirmed the orthogonality finding at a much larger scale, computing 16 interdisciplinarity measures for 224 Web of Science Subject Categories. Their coherence indicator — adapted from Wang (2016), who operationalized it as the number of citation links between cited references belonging to different categories weighted by their dissimilarity — showed only weak to moderate correlations with the diversity family: $r = 0.23$ with multi-assignment proportion, $r = 0.44$ with Simpson diversity, $r = 0.46$ with Shannon entropy, and $r = 0.50$ with the inverted Gini coefficient. The correlation with betweenness centrality was negligible ($r = -0.03$), and the correlation with the cluster coefficient was negative ($r = -0.36$). These results place coherence in a distinct empirical cluster from categorical diversity measures, reinforcing the claim that it captures a genuinely independent dimension of interdisciplinarity.

The diversity–coherence framework

The joint observation of diversity and coherence gives rise to a useful interpretive matrix, proposed as a two-dimensional framework by Rafols and Meyer (2009). Low diversity combined with high coherence characterizes specialized disciplinary research — tightly integrated work within a single paradigm. Low diversity with low coherence indicates that distant specialties within the same discipline are being connected, without yet achieving full integration. High diversity with low coherence represents the most nascent form of interdisciplinary integration: hitherto unrelated bodies of knowledge are being juxtaposed for the first time, as in the Noji (1997) case. Finally, high diversity with high coherence marks the mature state of specialized interdisciplinary research, where formerly distant knowledge sources have been woven into a stable intellectual fabric.

The framework implies a trajectory of knowledge integration that moves from low to high coherence over time: pioneering integration gradually consolidates into established interdisciplinary specialties. For evaluation purposes, this trajectory enables a distinction between early-stage integration (high potential, low consolidation) and mature interdisciplinary fields (high potential realized), a

nuance that scalar diversity measures alone cannot capture.

Alternative operationalizations

Betweenness centrality. Leydesdorff and Rafols (2011) explored betweenness centrality as an alternative coherence-related indicator at the journal level. Freeman's betweenness centrality, defined as

$$g_i = \sum_{\substack{j,k \\ j \neq k \neq i}} \frac{g_{ijk}}{g_{jk}}$$

where g_{jk} is the total number of geodesics between nodes j and k and g_{ijk} the number of those geodesics passing through i, measures the extent to which a journal occupies an intermediary position in the citation network. However, raw betweenness is confounded by size: large multi-disciplinary journals such as *Nature* and *Science* score highly simply because of their degree centrality. Leydesdorff and Rafols addressed this by computing betweenness in cosine-normalized networks, after which social-science journals emerged as the most prominent interdisciplinary bridges across 8,207 JCR journals. In a rotated factor analysis (Bollen et al., 2009), betweenness loaded near the origin — almost orthogonal to both citation-based and vector-based indicators — suggesting that it captures a distinct positional dimension.

Distance-measure sensitivity. Leydesdorff and Rafols (2011) also documented a striking sensitivity of Rao-Stirling diversity to the choice of distance matrix. When they computed the indicator using $(1 - \cos)$ versus relative Euclidean distances across the same 8,207 journals, the Spearman rank-order correlation between the two resulting interdisciplinarity rankings was $\rho = -0.012$ in the cited direction and $\rho = -0.015$ in the citing direction — effectively zero and, in the latter case, nominally negative. In a rotated factor analysis the Euclidean-based variant loaded on a different component from all other indicators, confirming that the two distance formulations capture fundamentally different structural features of the citation network. This dramatic finding underscores that the choice of distance measure is not a minor technical detail but a first-order determinant of measured interdisciplinarity.

IKF homogeneity. Zhou, Guns, and Engels (2023) proposed the Interdisciplinary Knowledge Flow (IKF) framework, which decomposes inter-field citation relationships into three aspects: *broadness* (the fraction of publications that cite a given external discipline), *intensity* (the share of outward citations directed at that discipline), and *homogeneity* (the fraction of a discipline's references that are co-cited by the target discipline). The homogeneity dimension is the most directly related to Rafols and Meyer's coherence concept: it measures cognitive similarity via the overlap of knowledge bases. Empirically, homogeneity correlates moderately with broadness ($R^2 = 0.47$) but only weakly with intensity ($R^2 = 0.25$), indicating that these aspects capture different facets of

the interdisciplinary relationship. A revealing pattern is that low homogeneity combined with high broadness characterizes methodological disciplines (e.g., Applied Mathematics cited by Ecology or Genetics): their tools diffuse broadly despite a large cognitive distance from the receiving field.

Computational considerations

All coherence indicators require a pairwise similarity or coupling computation whose cost grows as $O(N^2)$ in the number of publications or journals. For betweenness centrality the cost is higher, at $O(N^3)$ in the naive implementation, because shortest-path enumeration is required on the full network. In practice, Salton's cosine is preferred for constructing the coupling matrix because it is non-parametric, handles sparse reference vectors naturally, and normalizes for publication length. The co-occurrence matrix itself can be obtained efficiently via matrix multiplication ($\mathbf{A}\mathbf{A}^\top$ for bibliographic coupling, $\mathbf{A}^\top\mathbf{A}$ for co-citation). Threshold selection when binarizing valued networks remains a practical trade-off between noise filtering and information loss; the common choice of $\tau = 0.05$ is adequate for moderately sized reference lists but may need adjustment for fields with substantially different citation practices.

Summary

Coherence indicators complement diversity by measuring the depth of knowledge integration rather than the breadth of categorical spread. Multiple operationalizations are available — bibliographic coupling density, betweenness centrality, and co-citation homogeneity — and the empirical evidence consistently shows that they are only weakly correlated with diversity measures (typically $r < 0.5$), confirming that coherence constitutes an independent measurement dimension. Combined with diversity, coherence enables a richer characterization that distinguishes potential from realized integration, a distinction the next subsection builds upon as it examines indicators that explicitly combine both dimensions.

Diffusion Indicators

Conceptual foundations

The indicators examined in the preceding subsections — diversity, coherence, and their composites — all look *backward* from a publication to the knowledge it draws upon. Diffusion indicators invert the causal direction: they look *forward* from a publication to the new bodies of research that cite it, measuring the cross-disciplinary reach of knowledge outputs rather than the heterogeneity of knowledge inputs (Cantone, 2024; Leydesdorff, Wagner, and Bornmann, 2019). Formally, the reference vector $\mathbf{p}(x)$ that underlies integration measures is replaced by a citation vector $\mathbf{q}(x)$, where $q_i(x)$ denotes the proportion of *citing* publications that belong to disciplinary category i.

Cantone (2024) situates diffusion at the end of a temporal causal chain: cognition

precedes production, and diffusion follows both. Whereas integration captures the disciplinary breadth of inputs that shaped a piece of research, diffusion captures the disciplinary breadth of the communities that subsequently absorbed it. This asymmetry is consequential for evaluation. A study that integrates knowledge from many fields may nevertheless remain confined to its home discipline in terms of readership; conversely, a narrowly based study may diffuse widely if its methods or findings prove transferable. Diffusion therefore constitutes an independent measurement dimension, one that complements integration rather than duplicating it.

Operational definitions

The simplest operationalization of diffusion is the *fraction of citations received from outside the primary field*, analogous to the proportion of external references (p_{outside}) used for integration. More informative measures apply the same diversity machinery introduced in Section 3.1 but to the citation vector $\mathbf{q}(x)$ instead of the reference vector $\mathbf{p}(x)$.

Leydesdorff, Wagner, and Bornmann (2019) propose the *DIV* indicator, which decomposes diversity into its three Stirling components and applies them to the *cited* direction. For a journal c,

$$\text{DIV}_c = \frac{n_c}{N}\,(1 - G_c)\,\frac{\sum_{i \neq j} d_{ij}}{n_c(n_c - 1)}$$

where n_c is the number of Web of Science categories with non-zero citation shares (variety), N the total number of categories available, $d_{ij} = 1 - \cos(\mathbf{v}_i, \mathbf{v}_j)$ the disparity between categories i and j, and G_c the Gini coefficient measuring the unevenness of the citation distribution across categories. The balance component enters as $(1-G_c)$: perfect evenness yields $G_c = 0$ and maximum balance, whereas concentration in a single category yields $G_c \to 1$ and vanishing balance. Gini is computed via the ascending-order formula,

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1)\, x_i}{n \sum_{i=1}^{n} x_i}$$

where the x_i are the citation shares sorted in non-decreasing order. The three-way factorization makes DIV *monotonically* increasing in each component, a property that the Rao-Stirling index lacks because RS combines variety and balance *ex ante* through the Simpson concentration index (Leydesdorff, Wagner, and Bornmann, 2019).

A complementary measure is the *coherence* of the citing distribution,

$$C = \sum_{i \neq j} p_{ij}\, d_{ij}$$

which captures the average cognitive distance among co-occurring citation categories. When applied to the cited direction, high coherence indicates that the citing communities themselves span distant parts of the disciplinary landscape — a strong signal of diffusion breadth.

Temporal dynamics and measurement challenges

Unlike reference-based indicators, which are fixed at publication time, diffusion measures are inherently dynamic. Citations accumulate over months and years, so a diffusion score computed one year after publication may differ substantially from one computed five years later. This creates a time-window dependency that integration measures do not face. Furthermore, citations are not under author control: the same publication may attract citations from unexpected fields depending on shifts in research fashion, policy relevance, or methodological uptake (Cantone, 2024).

Three further complications arise. First, diffusion is confounded with *scientific impact*: highly cited publications receive citations from more categories simply by virtue of their citation volume, even if each individual citation comes from within the home discipline. Any diffusion measure must therefore be interpreted alongside total citation counts. Second, a non-trivial fraction of publications receive zero or near-zero citations within typical evaluation windows, rendering diffusion scores undefined or degenerate — a censoring problem with no clean analogue on the reference side. Third, citation practices vary across disciplines in both volume and latency, meaning that raw diffusion scores are not directly comparable across fields without normalization.

These challenges make diffusion indicators most appropriate for longitudinal or time-series analyses, where the temporal evolution of cross-disciplinary reach can be tracked explicitly rather than frozen at an arbitrary cutoff.

Empirical patterns from large-scale studies

Leydesdorff, Wagner, and Bornmann (2019) computed DIV in both the citing and cited directions for 11,487 journals indexed in the *Journal Citation Reports* (JCR 2016). In the cited direction (diffusion), *PLOS ONE* achieved the highest DIV score (0.142), followed by *Science* (0.125) and *Nature* (0.124) — journals whose editorial scope is broad enough to attract citing communities from across the disciplinary map. Notably, the Rao-Stirling diversity index applied to the same data produced a substantially different ranking: *Daedalus-US* (0.939), *Qualitative Inquiry* (0.936), and *Critical Inquiry* (0.927) led, illustrating how

measure choice can alter empirical conclusions even when the underlying data are identical.

The divergence is instructive. RS is dominated by its disparity term, which favours journals cited by a few very distant categories; DIV gives proportional weight to variety and balance, rewarding journals that attract citations from many categories in roughly even shares. In a factor analysis across the full journal set, DIV in the cited direction loaded on the same factor as betweenness centrality ($\rho = 0.66$) and impact factor, whereas RS loaded separately ($\rho = 0.41$ with betweenness). This pattern suggests that DIV captures a structural *intermediation* role — journals that serve as conduits between disciplinary communities — while RS captures a different aspect of cross-field reach that is less aligned with network position.

Distinction from topic-based interdisciplinarity

Diffusion is sometimes conflated with a related but distinct concept: *topic interdisciplinarity*, defined as the degree to which a publication's content addresses themes from multiple disciplines. Xiang, Romero, and Teplitskiy (2025) disentangle two dimensions that are often confounded in empirical work. *Topic interdisciplinarity* is measured through the disciplinary classification of a publication's title and abstract (e.g., OpenAlex concept tags), and reflects what the work *addresses*. *Knowledge-base interdisciplinarity* is measured through the disciplinary composition of the reference list, and reflects what the work *draws upon*. The two correlate only moderately ($r = 0.56$), confirming that they are not interchangeable.

Critically, the two dimensions carry opposite associations with peer review outcomes. In an analysis of 128,950 STEM manuscripts submitted between 2018 and 2022, Xiang, Romero, and Teplitskiy (2025) find that a one-standard-deviation increase in knowledge-base interdisciplinarity raises acceptance probability by 0.9 percentage points, while the same increase in topic interdisciplinarity *lowers* it by 1.2 percentage points. The interaction term is positive and significant ($\beta = 0.042$): broad references mitigate the penalty incurred by interdisciplinary framing. These findings underscore that the *direction* of disciplinary boundary-crossing matters: integration (references) is rewarded; topical spanning (content) is penalized unless supported by demonstrably broad knowledge inputs.

Relation to integration measures

Because diffusion and integration use the same mathematical machinery — Stirling diversity, Gini balance, disparity matrices — applied to different input vectors, a natural question is whether they are empirically related. Leydesdorff, Wagner, and Bornmann (2019) report that the correlation between DIV in the citing direction (integration) and DIV in the cited direction (diffusion) is positive but far from unity, confirming that the two capture different phenomena. High integration does not guarantee high diffusion, and vice versa.

Cantone (2024) speculates that integrative work may serve as an *antecedent macro-factor* for diffusion: research that synthesizes knowledge from multiple fields may become cognitively accessible to a broader audience, creating a "superspreader" effect. This hypothesis remains untested in the literature and would require longitudinal designs linking integration scores at publication time to diffusion trajectories in subsequent years.

Validity concerns and conceptual status

The conceptual status of diffusion within an interdisciplinarity measurement framework is contested. Cantone (2024) argues that diffusion "defies the full definition of IDR" because it measures a *future effect* of research rather than a property of the research *process*. Under this view, diffusion is an *outcome* of scientific activity — akin to impact — rather than an *attribute* of the activity itself. Whether a publication is cited by distant disciplines depends on factors largely external to the authors' integrative effort: editorial policies of citing journals, the availability of the work in relevant databases, and broader trends in research policy.

This conceptual tension admits two resolutions. One may treat diffusion as a legitimate dimension of interdisciplinarity, on the grounds that cross-disciplinary reach is itself a form of boundary-crossing that evaluation frameworks should capture. Alternatively, one may treat diffusion as a *consequence* of interdisciplinarity that is informative for policy but should not be conflated with the measurement of interdisciplinary production. The distinction is not merely semantic: it determines whether diffusion indicators belong in a panel designed to characterize research *as it is produced* or in a separate assessment of research *as it is received*. Wang and Schneider (2020) note that different indicators may capture "different understandings of such a multi-faceted concept as interdisciplinarity," a warning that applies with particular force to the integration–diffusion boundary.

Practical guidance and research gaps

Given the challenges outlined above, diffusion indicators are best suited to retrospective impact assessment and longitudinal evaluation rather than *ex ante* project selection or funding decisions. They are most informative when applied with explicit time windows (e.g., five-year or ten-year citation windows) and when accompanied by total citation counts that allow analysts to distinguish genuine cross-disciplinary reach from the mechanical effect of high citation volume.

Several research gaps remain. Citation-side Gini coefficients and DIV scores are not routinely computed in standard bibliometric toolkits, limiting their practical uptake. The temporal structure of diffusion — whether cross-disciplinary citing patterns stabilize or continue to evolve decades after publication — has received little systematic attention. Finally, the hypothesized feedback loop between integration and diffusion — whereby broadly integrative work attracts broader citing audiences, which in turn stimulates further integration — remains an open

empirical question that longitudinal panel designs are well positioned to address.

Novelty Indicators

Conceptual foundation

Novelty indicators address a question distinct from diversity: whether research combines disciplinary elements in ways that are not merely heterogeneous but genuinely *nonconforming* relative to established practice. A publication may exhibit high diversity — drawing on many distant fields — yet low novelty if its particular combination of fields has become routine, effectively constituting a new "interdisciplinary discipline" (Cantone, 2024). Novelty captures disciplinary *innovation*, not disciplinary *breadth*.

The conceptual core is divergence from a benchmark. Given a disciplinary proportion vector $p(x)$ for a research unit x and a benchmark distribution $p(E)$ representing expected disciplinary composition, novelty is operationalized as a function of the discrepancy between the two:

$$\nabla = f\big(|p(x) - p(E)|\big).$$

The choice of benchmark, the functional form f, and the treatment of disciplinary similarity $z(i, j)$ distinguish the competing approaches reviewed below. A revealing limiting case connects novelty to diversity: when $p(x)$ and $p(E)$ have non-overlapping support, the squared divergence decomposes as $\sum_i [p_i(x) - p_i(E)]^2 = \sum_i p_i(x)^2 + \sum_i p_i(E)^2$, where each summand is a rate-of-repeat (Herfindahl) term — a diversity measure (Cantone, 2024). Divergence thus generalizes diversity by incorporating prior expectations.

Statistical divergence approaches

Several families of divergence measures have been proposed for novelty quantification. The simplest is the sum of squared differences between $p(x)$ and $p(E)$, but it has seen no empirical adoption. Chi-squared divergence and related information-theoretic measures (Kullback–Leibler, mutual information) suffer from a common flaw: they are undefined when $p_i(E) = 0$ for any category i where $p_i(x) > 0$, forcing analysts to ignore precisely the most novel disciplinary contributions (Cantone, 2024).

Two alternatives avoid this singularity. The *probabilistic Jaccard* index compares distributions through their overlap:

$$\nabla_{\mathrm{PJ}}(x) = 1 - \frac{\sum_i \min\big(p_i(x),\, p_i(E)\big)}{\sum_i \max\big(p_i(x),\, p_i(E)\big)}.$$

This is a normalized variant of generalized mutual entropy with a purely frequentist interpretation (Moulton and Jiang, 2018). It is easy to compute but difficult

to interpret in substantive terms (Cantone, 2024). The *Hellinger distance* takes a geometric approach:

$$\nabla_{\text{Hel}}(x) = \frac{1}{\sqrt{2}} \sqrt{\sum_i \Big(\sqrt{p_i(x)} - \sqrt{p_i(E)} \Big)^2}.$$

Hellinger distance offers stable scalings and well-understood metric properties but depends on a Euclidean interpretation of probability space that is difficult to communicate to non-specialist audiences (Cantone, 2024).

Neither the probabilistic Jaccard nor the Hellinger distance accounts for inter-category similarity $z(i, j)$. A similarity-weighted extension replaces each proportion p_i with a smoothed version $\psi_{i,z}(x) = [\sum_j p_j(x)\, z(i, j)]/[\sum_i \sum_j p_j(x)\, z(i, j)]$, which can then be substituted into any divergence formula (Cantone, 2024). This extension brings novelty measurement into closer alignment with disparity-aware diversity indices, but at the cost of requiring the same similarity matrix infrastructure.

Permutation method

The most influential novelty measure in the bibliometric literature is the atypical-combinations approach of Uzzi et al. (2013). Rather than comparing aggregate distributions, it examines all pairwise co-occurrences of disciplinary categories within a body of research. For each observed pair (i, j), a z-score is computed against a randomized null model:

$$\nabla_{\text{Uzzi}}(p_{i,j},\, x) = \frac{p_{i,j}(x) - e_{i,j}(x)}{\sigma\big[e_{i,j}(x)\big]}, \quad p_{i,j}(x) > 0,$$

where $e_{i,j}(x)$ and $\sigma[e_{i,j}(x)]$ are the mean and standard deviation of the pair frequency under permuted citation networks that preserve reference counts. The method produces a *distribution* of atypicality scores within x, from which Uzzi et al. derive two non-parametric summary statistics capturing conformity and novelty.

The empirical claims that emerged from this framework were highly influential for research policy. Uzzi et al. (2013) reported that the highest-impact science is "primarily grounded in exceptionally conventional combinations" yet simultaneously features "intrusion of unusual combinations," and that teams are more likely than solo authors to insert novel pairings into familiar knowledge domains.

Subsequent validation work, however, has cast serious doubt on whether the Uzzi measure captures novelty as a construct distinct from diversity. Bornmann (2019), using expert assessments from F1000Prime as ground truth, found that the atypical-combinations measure did *not* correlate with qualitatively assessed novelty. Fontana et al. (2020) confirmed this finding and showed that

the Uzzi measure instead correlates with Rao–Stirling diversity — precisely the construct it was designed to complement, not replicate. This conflation of novelty and diversity within a single indicator undermines the original interpretive framework and calls into question policy conclusions drawn from it. Cantone (2024) identifies a structural reason for the conflation: the permutation method conflates the novelty and diversity dimensions into a single measure rather than separating them systematically. In addition, the method imposes very high data requirements, needing complete reference lists and sufficient network density for reliable null-model estimation.

Timed novelty

Wang et al. (2017) propose a temporal approach that lies conceptually between statistical nonconformity and trailblazing: for each pair (i, j) of disciplinary categories with $p_{i,j} > 0$, they record the timestamp $t_0(i, j)$ of its first observed occurrence in a reference corpus. The timed novelty score is then:

$$\text{Novelty}(x) = \sum_{i,j} t_0(i,j) \cdot \big[1 - z(i,j)\big],$$

weighting recency of first combination by inter-category disparity. The measure is intuitive — it rewards research that instantiates disciplinary pairings not previously observed, especially between cognitively distant fields.

However, timed novelty faces multiple serious limitations. It requires historical citation data sufficient to identify first occurrences reliably; in samples containing only recent publications, all pairs appear novel by construction, producing misleading scores (Cantone, 2024). The measure is also highly sensitive to taxonomy granularity, rendering it inappropriate for coarse-grained classification systems. Most critically, the measure lacks a mechanism for calibrating innovation against diffusion: a "pioneering" combination that no subsequent work ever cites raises the question of whether it constitutes genuine innovation or merely an unproductive anomaly. Bornmann (2019) and Fontana et al. (2020) found no concordance between timed novelty scores and expert-assessed novelty, questioning the measure's epistemic validity.

Benchmark specification

All divergence-based novelty measures depend on the choice of benchmark $p(E)$, and this choice is far from neutral. Three approaches have been proposed. Goyanes et al. (2020) adopt a *uniform prior* (equal weight across all observed categories), which Cantone (2024) criticizes as unrealistic: "virtually no real applications expect a perfectly balanced distribution." Uzzi et al. (2013) use *randomized permutation*, swapping citations while preserving reference counts to generate a data-driven null model — the preferred approach for citational analyses. Cantone and Nightingale (2024) propose a *hierarchical benchmark* in

which the disciplinary distribution of a containing unit (e.g., a journal) serves as $p(E)$ for its constituent papers, exploiting the natural nesting of publication units. Each approach carries assumptions about what constitutes "expected" disciplinary composition, and no consensus has emerged on best practice.

Exclusion from the panel

Despite their conceptual importance, novelty indicators are excluded from the measurement panel proposed in this paper. The reasons are both operational and epistemic. On the operational side, all viable approaches require data infrastructure — complete citation networks, historical temporal baselines, or reference-list permutation apparatus — that substantially exceeds what standard bibliometric toolkits provide. Permutation methods impose "very high requirements" (Cantone, 2024), and network-based metrics are rarely unbiased at typical sample sizes. On the epistemic side, the two most prominent approaches — the Uzzi permutation method and Wang timed novelty — have both failed external validation, showing no concordance with expert-assessed novelty (Bornmann, 2019; Fontana et al., 2020) and correlating instead with diversity constructs the measures were designed to distinguish from. The absence of novelty measures from the 23-indicator review of Wang and Schneider (2020) further attests to their limited integration into standard practice. Novelty remains an important conceptual dimension of interdisciplinarity, and future work on validated operationalizations may warrant its inclusion; for now, the panel focuses on dimensions — diversity, coherence, and diffusion — for which measurement tools have stronger empirical grounding.

Mathematical Coverage and Qualification Map

To make the mathematical scope of the reviewed indicators explicit, the table below maps the principal formula-bearing references used in this review to their qualification conditions. This is a claim-hygiene device: each formula family is tied to at least one boundary condition that constrains interpretation.

Formula family	Primary references	Qualification condition used in this review
Rao-Stirling / integration (`\Delta = \sum d_{ij} p_i p_j`)	Porter and Rafols (2009); Rafols and Meyer (2009); Leydesdorff and Rafols (2011)	Values are not invariant to taxonomy granularity, similarity-matrix construction, or distance metric choice.
Similarity-based true diversity (`{}^qD^S`, `{}^2D^S`)	Zhang, Rousseau, and Glanzel (2016); Hill (1973); Jost (2006, 2009); Leinster and Cobbold (2012)	Entropy-like quantities must be interpreted on an effective-number scale; similarity-based and disparity-based variants are not numerically interchangeable.

Formula family	Primary references	Qualification condition used in this review
Variety-balance indices (Shannon, Simpson, Herfindahl, Gini)	Porter and Rafols (2009); Leydesdorff, Wagner, and Bornmann (2019); Mutz (2022)	These indices cannot, by themselves, identify cognitive distance; aggregation operator choice (additive vs multiplicative) changes rankings.
Coherence (`S`) via bibliographic coupling	Rafols and Meyer (2009); Jensen and Lutkouskaya (2014)	Coherence estimates depend on coupling thresholding/binarization and network-construction conventions.
Centrality-based alternatives	Leydesdorff and Rafols (2011); Bollen et al. (2009)	Betweenness and related graph indicators mix interdisciplinarity with size/position effects unless normalized carefully.
Diffusion / cross-field effect (`E`)	Leydesdorff, Wagner, and Bornmann (2019); Xiang, Romero, and Teplitskiy (2025); Lariviere and Gingras (2010)	Cross-field uptake must be field-normalized; diffusion should be treated as distinct from input diversity.
Knowledge-flow decomposition (`B, I, H`)	Zhou, Guns, and Engels (2023)	Distributional flow vectors answer directional exchange questions but are not drop-in replacements for scalar panel components.
Novelty via atypical combinations / timed emergence	Uzzi et al. (2013); Wang et al. (2017); Bornmann (2019); Fontana et al. (2020)	Novelty indicators capture atypicality under explicit null-model assumptions; external validation remains limited for policy use.
Near-zero overlap stabilization	Moulton and Jiang (2018)	Probabilistic Jaccard variants improve zero-overlap behavior but still require complementary disparity modeling for interdisciplinarity claims.

Formula family	Primary references	Qualification condition used in this review
Uncertainty quantification and validity diagnostics	Zwanenburg, Nakhoda, and Whigham (2022); Nakhoda, Whigham, and Zwanenburg (2023)	Individual-level estimates require interval reporting; point-estimate thresholding alone is decision-fragile.
Empirical coherence benchmark corpus	Noji et al. (1997); Tomishige et al. (2002)	These are benchmark test cases for indicator behavior, not normative formulas for interdisciplinarity quality.
Transdisciplinary quality framing	Stokols et al. (2003); Klein (2008); Borlaug and Svartefoss (2025)	Quality judgments require explicit evaluative criteria beyond bibliometric panel values.

This map does not claim that every cited paper contributes a novel formula. Rather, it makes explicit how the mathematics that *is* used in the review is qualified before being translated into evaluation guidance.

Beyond Scalars: Distribution-Based Approaches

Recent work has challenged the assumption that interdisciplinarity should be measured by scalar indicators at all. Zhou, Guns, and Engels (2023) propose an Interdisciplinary Knowledge Flow (IKF) framework that characterizes the relationship between any two disciplines along three aspects: *broadness* (what fraction of publications cite a given external discipline), *intensity* (how deeply engaged those citing publications are), and *homogeneity* (cognitive similarity via co-citation overlap). Formally, given a citation matrix M $(n \times n)$ and entities X (citing) and Y (cited), broadness is $B(X,Y) = |X'|/|X|$, where X' is the subset of publications in X that cite at least one publication in Y. Intensity restricts the denominator to outward citations from X' only: $I(X,Y) = \sum_{i \in X, j \in Y} M_{ij} / \sum_{i \in X, j=1}^{n} (M_{ij}\, \delta_i)$, where $\delta_i = 1$ iff $i \in X'$. Homogeneity measures knowledge-base overlap: $H(X,Y) = \sum_{i \in X, \gamma=1}^{n} M_{i\gamma}\, \varphi_{\gamma,Y} / \sum_{i \in X, j=1}^{n} M_{ij}$, where $\varphi_{\gamma,Y} = 1$ if publication γ is also cited by Y. Each aspect is thus a well-defined fraction, and the triple (B, I, H) jointly characterizes the *form* of interdisciplinary knowledge exchange — yielding a distribution vector rather than a single number and answering "what is interdisciplined" rather than merely "how interdisciplinary."

Cantone (2024) takes a complementary systemic approach, decomposing the measurement problem into a pipeline of analytical choices: selection of the unit of analysis (paper, author, institution), choice of disciplinary taxonomy, classifi-

cation method, operational definition (dimension and formula), and aggregation strategy. This framing makes explicit that indicator values depend on a chain of methodological decisions, each of which introduces potential inconsistency.

Both approaches suggest that the field is moving away from single-number summaries toward richer, multidimensional characterizations. Our panel occupies a middle ground: it is multidimensional (three components) but produces a compact, interpretable profile rather than a high-dimensional distribution.

Taxonomy Summary

The following table organizes major indicators by conceptual dimension and methodological family:

Dimension	Reference-based	Citation-based	Text-based	Network-based
Diversity	Rao-Stirling, Simpson, Shannon, Gini, Hill	—	Semantic similarity	—
Coherence	Bibliographic coupling density	—	—	Betweenness centrality, clustering
Diffusion	—	Cross-field citations, citing diversity	—	—
Novelty	—	—	—	Recombination metrics

Four observations emerge. First, the reference-based/diversity cell is heavily populated while other cells remain sparse — the literature has focused disproportionately on measuring diversity of inputs. Second, coherence and diffusion are largely orthogonal to diversity (confirmed empirically by Wang and Schneider, 2020), yet receive far less attention. Third, text-based and network-based methods are underrepresented relative to their potential. Fourth, no existing indicator spans multiple dimensions, motivating the multi-component approach we develop next.

A Multi-Component Panel

Against the backdrop of the indicator landscape surveyed above, we now present a specific three-component panel designed to span three of the four identified dimensions: diversity, coherence, and diffusion.

Panel Definition

We define three indicators that together characterize the interdisciplinarity of a researcher's output portfolio. Each captures a distinct dimension of knowledge integration.

Rao-Stirling Diversity

Let p_i denote the proportion of references in category i across a researcher's publications, and let s_{ij} denote the cosine similarity between categories i and j (computed from aggregate citation patterns). Define the distance $d_{ij} = 1 - s_{ij}$. The Rao-Stirling diversity index is

$$\Delta = \sum_{i,j} d_{ij}\, p_i\, p_j = 1 - \sum_{i,j} s_{ij}\, p_i\, p_j$$

This is the variant with $\alpha = \beta = 1$ of Stirling's (2007) generalized diversity heuristic. It reduces to the Simpson diversity index when all categories are maximally disparate ($s_{ij} = 0$ for $i \neq j$), and it equals zero when all references fall in a single category. The index captures variety, balance, and disparity simultaneously (Porter and Rafols, 2009).

Network Coherence: Mean Linkage Strength

Let a researcher have n publications, and let $\mathbf{r}_k$ be the reference vector of publication k over the set of categories. We define the mean linkage strength as

$$S = \frac{1}{\binom{n}{2}} \sum_{k<l} \cos(\mathbf{r}_k, \mathbf{r}_l)$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. This is a bibliographic coupling measure: publications that share many references in similar categories have high pairwise similarity. A high value of S indicates that the researcher's publications form a coherent body of work; a low value indicates disconnected contributions across unrelated topics.

The coherence indicator was introduced conceptually by Rafols and Meyer (2009), who operationalized it as the mean density of bibliographic coupling networks. It provides a bottom-up perspective that complements the top-down diversity measure.

Cross-Field Effect Proxy

For each of a researcher's publications, let k^* denote its primary category (the category with the largest share of references). The cross-field effect proxy is

$$E = \frac{\text{citations from articles whose primary category} \neq k^*}{\text{total citations received}}$$

where the sum is pooled across all of a researcher's publications. A high value of E indicates that the researcher's work is used across disciplinary boundaries — it produces cross-field impact, not merely cross-field inputs. This distinguishes genuine integration (high Δ, moderate S, high E) from polymathic breadth (high Δ, low S, low E).

Discrimination on Toy Data

We illustrate the panel with a toy university dataset consisting of five subject categories and three researcher archetypes.

Setup

The five categories are: condensed matter physics (C_1), materials science (C_2), physical chemistry (C_3), molecular biology (C_4), and applied mathematics (C_5). Their pairwise similarities are given by the matrix

	C_1	C_2	C_3	C_4	C_5
C_1	1.00	0.60	0.40	0.10	0.30
C_2	0.60	1.00	0.50	0.15	0.20
C_3	0.40	0.50	1.00	0.35	0.10
C_4	0.10	0.15	0.35	1.00	0.05
C_5	0.30	0.20	0.10	0.05	1.00

Each researcher has five publications with reference vectors over these categories. The aggregate category proportions p_i are derived from these vectors (not specified independently), ensuring internal consistency across all three indicators.

- **Researcher A** (cross-disciplinary integrator): Each publication references multiple distant categories (e.g., condensed matter and molecular biology in a single paper). Aggregate: $p_A = (0.300, 0.075, 0.175, 0.350, 0.100)$.
- **Researcher B** (polymath): Five single-field publications, one per category. Aggregate: $p_B = (0.200, 0.200, 0.200, 0.200, 0.200)$.
- **Researcher C** (specialist): All publications concentrated in condensed matter and materials science. Aggregate: $p_C = (0.636, 0.273, 0.061, 0.000, 0.030)$.

Worked Computational Example

To make the panel computation fully transparent and reproducible, we trace the calculation of all three indicators for each researcher from the raw publication-level reference vectors. The complete data are presented below; the proportions p_i reported above are derived from these vectors, not specified independently.

Publication reference vectors. Each entry $r_{k,i}$ gives the number of references from publication k to category C_i:

Publication	C_1	C_2	C_3	C_4	C_5	Total
A1	3	0	2	3	0	8
A2	4	1	0	3	0	8
A3	2	2	0	3	1	8
A4	0	0	3	3	2	8
A5	3	0	2	2	1	8
A total	**12**	**3**	**7**	**14**	**4**	**40**
B1	5	0	0	0	0	5
B2	0	5	0	0	0	5
B3	0	0	5	0	0	5
B4	0	0	0	5	0	5
B5	0	0	0	0	5	5
B total	**5**	**5**	**5**	**5**	**5**	**25**
C1p	4	3	0	0	0	7
C2p	5	1	0	0	0	6
C3p	3	3	1	0	0	7
C4p	4	2	1	0	0	7
C5p	5	0	0	0	1	6
C total	**21**	**9**	**2**	**0**	**1**	**33**

Step 1: Diversity (Δ). The computation proceeds in three stages. First, derive aggregate proportions: for Researcher A, $p_1 = 12/40 = 0.300$, $p_2 = 3/40 = 0.075$, $p_3 = 7/40 = 0.175$, $p_4 = 14/40 = 0.350$, $p_5 = 4/40 = 0.100$. Second, compute the Herfindahl concentration index $H = \sum_i p_i^2$:

$$H_A = 0.300^2 + 0.075^2 + 0.175^2 + 0.350^2 + 0.100^2 = 0.258750$$

Third, compute the full similarity-weighted sum $\sum_{i,j} s_{ij}\, p_i\, p_j = H + 2\sum_{i<j} s_{ij}\, p_i\, p_j$. The dominant cross-terms for Researcher A are:

Pair (i, j)	s_{ij}	p_i	p_j	$s_{ij}\, p_i\, p_j$
(C_1, C_4)	0.10	0.300	0.350	0.010500
(C_1, C_3)	0.40	0.300	0.175	0.021000
(C_3, C_4)	0.35	0.175	0.350	0.021438
(C_1, C_2)	0.60	0.300	0.075	0.013500
(C_1, C_5)	0.30	0.300	0.100	0.009000

Summing all ten off-diagonal pairs and applying the symmetry factor yields $\sum_{i,j} s_{ij}\, p_i\, p_j = 0.440625$, whence $\Delta_A = 1 - 0.440625 = 0.559375$.

For Researcher B, the uniform distribution $p_i = 0.200$ gives $H_B = 0.200$ and $\sum_{i,j} s_{ij}\, p_i\, p_j = 0.420$, so $\Delta_B = 0.580$. For Researcher C, the concentrated distribution gives $H_C = 0.4839$ and $\Delta_C = 0.245$.

Step 2: Coherence (S). We compute all $\binom{5}{2} = 10$ pairwise cosine similarities between publication reference vectors. For Researcher A, representative values include $\cos(\mathbf{r}_{A1}, \mathbf{r}_{A2}) = 0.878$ (A1 and A2 both reference C_1 and C_4, sharing a physics–biology bridge) and $\cos(\mathbf{r}_{A2}, \mathbf{r}_{A4}) = 0.376$ (the weakest pair, as A4 avoids C_1 entirely). The mean over all ten pairs is $S_A = 7.334/10 = 0.733$.

For Researcher B, every publication vector is orthogonal to every other (each references exactly one category, and no two share a category), so $\cos(\mathbf{r}_{Bk}, \mathbf{r}_{Bl}) = 0$ for all $k \neq l$, giving $S_B = 0.000$. This zero coherence is diagnostic of polymathy: maximal breadth with no integration between publications.

For Researcher C, all publications cluster in the C_1–C_2 neighborhood, producing uniformly high pairwise cosines (range 0.675 to 0.964) and $S_C = 0.881$.

Step 3: Cross-field effect (E). For each publication, we assign the primary category as $k^* = \arg\max_i r_{k,i}$ (ties broken by lowest index). The citation breakdown is:

Publication	Primary	Total cites	From primary	From other
A1	C_1	6	2	4
A2	C_1	5	2	3
A3	C_4	4	2	2
A4	C_3	5	2	3
A5	C_1	5	2	3
A total		**25**	**10**	**15**

Thus $E_A = 15/25 = 0.600$: sixty percent of Researcher A's citations originate outside the citing publication's primary category, confirming genuine cross-field impact. For Researcher B, nearly all citations come from within each publication's own field ($E_B = 1/16 = 0.063$); for Researcher C, a small fraction of citations arrive from neighboring fields ($E_C = 4/19 = 0.211$).

Results

The full panel values are:

Researcher	Δ	S	E	Type
A (integrator)	0.559	0.733	0.600	Cross-disciplinary
B (polymath)	0.580	0.000	0.063	Polymathic breadth
C (specialist)	0.245	0.881	0.211	Disciplinary

The critical observation is that $\Delta_A \approx \Delta_B$ (0.559 versus 0.580): diversity alone cannot distinguish the integrator from the polymath. Both researchers draw on a broad range of categories, and the Rao-Stirling index correctly reports high diversity for both. The distinction lies in how that diversity is structured.

The coherence indicator S reveals the difference: Researcher A's publications share references across category boundaries ($S = 0.733$), while Researcher B's publications have no overlap at all ($S = 0$). The cross-field effect E confirms this at the impact level: Researcher A's work is cited across disciplines ($E = 0.600$), while Researcher B's single-field contributions are cited almost exclusively within their own fields ($E = 0.063$).

No single component of the panel achieves full discrimination. Diversity alone fails on A versus B. Coherence alone fails to distinguish A (moderate-high) from C (very high) without the context of diversity. The cross-field effect separates A from both B and C, but cannot on its own distinguish integrators from specialists when diversity is unknown. Only the full triple uniquely characterizes each type.

Sensitivity Analysis and Robustness

A natural concern is whether the panel's discrimination depends on the precise values of the inter-category similarity matrix s_{ij}. Because Δ is the only panel component that uses s_{ij}, robustness analysis centres on two questions: (i) how does Δ respond to perturbations of the similarity matrix, and (ii) are the other components affected at all? We address both analytically and through numerical experiments.

Analytical result: uniform perturbation formula

Proposition. Under a uniform additive perturbation $s_{ij} \to s_{ij} + \varepsilon$ for all $i \neq j$ (with diagonal entries unchanged):

$$\Delta_{\text{new}} = \Delta_{\text{old}} - \varepsilon\,(1 - H)$$

where $H = \sum_i p_i^2$ is the Herfindahl concentration index.

Proof. Write $\Delta = 1 - \sum_{i,j} s_{ij}\, p_i\, p_j$. Under the perturbation, the sum changes by $\varepsilon \sum_{i \neq j} p_i\, p_j$. Since $\sum_{i,j} p_i\, p_j = \left(\sum_i p_i\right)^2 = 1$ and $\sum_i p_i^2 = H$, the off-diagonal sum equals $1 - H$, giving the result. □

The formula makes the dependence on researcher concentration explicit: a more concentrated portfolio (higher H) experiences a smaller absolute shift in Δ for the same perturbation magnitude, because fewer distinct category pairs contribute to the off-diagonal sum.

Corollary (gap evolution). The signed gap between any two researchers evolves linearly:

$$\Delta_{A,\text{new}} - \Delta_{B,\text{new}} = (\Delta_A - \Delta_B) + \varepsilon\,(H_A - H_B)$$

This gap vanishes at the critical perturbation

$$\varepsilon^* = -\frac{\Delta_A - \Delta_B}{H_A - H_B}$$

at which point the diversity ranking of the two researchers inverts.

Critical inversion point

For the toy data, the exact Herfindahl indices are $H_A = 0.258750$ and $H_B = 0.200000$, giving $H_A - H_B = 0.058750$. The exact diversity gap is $\Delta_A - \Delta_B = 0.559375 - 0.580000 = -0.020625$. The critical inversion point is therefore

$$\varepsilon^* = \frac{0.020625}{0.058750} = 0.35106$$

This value requires that *every* off-diagonal similarity in the matrix be shifted by more than 0.35 — a perturbation exceeding 35% of the similarity scale — before the diversity ordering of Researchers A and B inverts. Since realistic uncertainty in citation-based similarity estimates is far smaller (Wang and Schneider, 2020, report inter-variant correlations of 0.30–0.91, corresponding to much smaller absolute shifts in individual s_{ij} entries), the near-equality $\Delta_A \approx \Delta_B$ is structurally robust rather than an artifact of the particular matrix chosen.

A technical note on computation: the exact value $\varepsilon^* = 0.35106$ must be derived from unrounded intermediate quantities. Using the rounded gap 0.021 and rounded Herfindahl difference 0.059 yields the approximation 0.356, a discrepancy of 1.4% that, while small, illustrates how rounding at intermediate stages can accumulate in derived quantities.

Gap evolution analysis

The corollary above implies that the gap $\Delta_A - \Delta_B$ evolves as a linear function of ε with slope $H_A - H_B = 0.058750 > 0$. Three regimes are distinguishable:

- For $\varepsilon < 0$ (categories become less similar): the gap widens in favour of B, but the absolute magnitude remains small.
- For $0 \leq \varepsilon < \varepsilon^*$: $\Delta_A < \Delta_B$, with the gap shrinking from its original value of 0.021 toward zero.
- For $\varepsilon > \varepsilon^*$: $\Delta_A > \Delta_B$, but the gap grows slowly (slope 0.059 per unit of ε).

Throughout this range, the separation between the high-diversity pair $\{A, B\}$ and the specialist C evolves as

$$\overline{\Delta}_{AB} - \Delta_C = \overline{\Delta}_{AB,0} - \Delta_{C,0} - \varepsilon \left(\overline{1 - H}_{AB} - (1 - H_C)\right)$$

where $\overline{\Delta}_{AB,0} = 0.5697$ and $\overline{1 - H}_{AB} = 0.7706$, $(1 - H_C) = 0.5161$. The coefficient on ε is -0.255, meaning the A/B–C separation decreases slowly as similarities increase but remains above 0.24 even at $\varepsilon = 0.35$. The panel's ability to separate specialists from broad researchers is preserved across all realistic perturbation magnitudes.

Non-uniform perturbation experiments

Uniform perturbation is a worst case in a precise sense: it shifts all similarities in the same direction, maximizing the cumulative effect on Δ. Real-world uncertainty in similarity matrices is more heterogeneous. We therefore tested two additional scenarios that model realistic patterns of matrix uncertainty.

Scenario 1: Neighbors closer. Adjacent categories (those with index distance $|i - j| = 1$) become more similar by 0.10, while distant categories ($|i - j| \geq 2$) become less similar by 0.05. This models a situation where fine-grained disciplinary boundaries become blurred while the macro-structure of knowledge is preserved.

Scenario 2: Uniform shift +0.10. All off-diagonal similarities increase by 0.10, corresponding to a citation database in which fields have become more interconnected (e.g., through the rise of data-driven methods applied across disciplines).

The results are summarized below:

Scenario	Δ_A	Δ_B	Δ_C	$\lvert\Delta_A - \Delta_B\rvert$	$\lvert\overline{\Delta}_{AB} - \Delta_C\rvert$
Original	0.559	0.580	0.245	0.021	0.32
Neighbors closer	0.557	0.572	0.214	0.015	0.35
Uniform +0.10	0.485	0.500	0.194	0.015	0.30

In both scenarios, three properties are preserved: (i) the A–B gap remains small (0.015, narrower than the original 0.021), confirming that diversity alone cannot separate integrators from polymaths regardless of matrix specification; (ii) the separation from C remains large (0.30 or above), ensuring clear identification of specialists; and (iii) the relative ordering $\Delta_C < \Delta_A < \Delta_B$ is maintained. The non-uniform perturbation (Scenario 1) actually *increases* the A/B–C separation because specialist portfolios, concentrated in neighboring categories, are more affected by neighbor-similarity changes than diverse portfolios.

Invariance of coherence and cross-field effect

A distinctive advantage of the multi-component approach is that only one of the three panel indicators depends on the similarity matrix. The coherence indicator S is computed from pairwise cosine similarities between publication reference vectors $\mathbf{r}_k$:

$$S = \frac{1}{\binom{n}{2}} \sum_{k<l} \frac{\mathbf{r}_k \cdot \mathbf{r}_l}{\|\mathbf{r}_k\| \, \|\mathbf{r}_l\|}$$

This quantity depends exclusively on the reference vectors themselves, not on any inter-category similarity structure. The cross-field effect E depends on citation flows and primary-category assignments (determined by $\arg\max_i r_{k,i}$), which are likewise independent of s_{ij}. Both S and E are therefore *exactly invariant* under any perturbation of the similarity matrix, whether uniform or non-uniform.

This invariance has a practical consequence: the discrimination between Researcher A ($S = 0.733$, $E = 0.600$) and Researcher B ($S = 0.000$, $E = 0.063$) is completely unaffected by the choice of similarity matrix. The multi-component panel is thus substantially more robust than any single-indicator approach based on diversity alone, because the coherence and cross-field effect channels carry no similarity-matrix uncertainty whatsoever.

Robustness summary

The sensitivity analysis yields three conclusions. First, the analytical perturbation formula $\Delta_{\text{new}} = \Delta_{\text{old}} - \varepsilon(1 - H)$ makes diversity shifts fully predictable: there are no threshold effects or nonlinear surprises below the inversion point ε^*. Second, the critical perturbation required to invert even the smallest diversity gap in our data ($\varepsilon^* = 0.351$) far exceeds realistic uncertainty in similarity estimation. Third, and most importantly, the coherence and cross-field effect indicators are completely immune to similarity-matrix perturbation, ensuring that the panel's discrimination power is preserved even when the diversity component is subject to specification uncertainty.

Interpretation Framework

The panel triple (Δ, S, E) is designed not merely as a measurement instrument but as a practical classification tool for evaluation contexts. This subsection provides a systematic framework for translating panel profiles into evaluation decisions.

Pattern taxonomy

Four recurring patterns emerge from the joint observation of the three panel components:

Pattern 1: High Δ, high S, high E — genuine integrator. The researcher draws on a broad range of disciplinary categories (high diversity), weaves them into a coherent body of work with substantial bibliographic coupling across publications (high coherence), and produces research that is cited across disciplinary boundaries (high cross-field effect). This is the canonical profile of cross-disciplinary integration. In the toy data, Researcher A exemplifies this pattern with $(\Delta, S, E) = (0.559, 0.733, 0.600)$.

Pattern 2: High Δ, low S, low E — polymath. The researcher publishes across many fields, producing high categorical diversity, but publications are mutually incoherent (low or zero bibliographic coupling) and each is cited primarily within its own field (low cross-field effect). This profile indicates breadth without integration — a collection of independent disciplinary contributions rather than a synthesized research programme. Researcher B exemplifies this pattern with $(\Delta, S, E) = (0.580, 0.000, 0.063)$.

Pattern 3: Low Δ, high S, low E — specialist. The researcher works within a narrow disciplinary cluster, producing low diversity but high internal coherence. Cross-field impact is limited because the work addresses a specialized audience. A researcher exhibiting this profile who has been classified as "interdisciplinary" by an evaluation agency is likely misclassified and should be redirected to standard disciplinary evaluation. Researcher C exemplifies this pattern with $(\Delta, S, E) = (0.245, 0.881, 0.211)$.

Pattern 4: Low Δ, low S, any E — emergent or insufficient data. When both diversity and coherence are low, the panel signals either an early-career researcher whose publication record is too sparse for stable estimation, or a researcher in an emerging field where disciplinary categories have not yet stabilized. In either case, the quantitative profile should be interpreted with caution, and qualitative expert assessment may be more appropriate.

Classification thresholds

For operational deployment, we propose illustrative decision thresholds:

Classification	Δ	S	E
Genuine integrator	≥ 0.40	≥ 0.30	≥ 0.30
Polymath (non-integrative)	≥ 0.40	< 0.15	< 0.15
Specialist (reclassify)	< 0.35	any	any
Ambiguous (expert review)	else	else	else

These thresholds are derived from the toy-data analysis and should be understood as starting points rather than universal cutoffs. Evaluation agencies must calibrate thresholds to their specific context, taking into account the granularity of the disciplinary classification system, the citation norms of the fields under review, and the policy objectives of the evaluation exercise. Validation on known

cases — researchers whose interdisciplinary status has been established through peer review or expert consensus — is essential before operational deployment.

Edge cases and failure modes

Six failure modes have been identified that may compromise panel-based classification, along with their recommended mitigations:

1. *Breadth-without-depth reward:* High Δ rewarded regardless of integration evidence. Mitigation: require S and E to exceed minimum thresholds before classifying as "integrator." This is precisely the discrimination the panel is designed to provide.
2. *Non-standard publication penalty:* Interdisciplinary journals often have lower impact factors. Mitigation: use field-normalized citation indicators; do not compare impact factors across disciplinary boundaries.
3. *Incommensurable citation norms:* Citation rates differ by an order of magnitude across fields (e.g., mathematics versus molecular biology). Mitigation: normalize E by field-specific citation baselines before cross-field comparison.
4. *Misclassification persistence:* A specialist enters an "Interdisciplinary" evaluation track and remains there indefinitely. Mitigation: automatic reclassification trigger when $\Delta < 0.35$.
5. *Early-career data sparsity:* Researchers with fewer than approximately 15 publications produce unstable panel estimates. Mitigation: impose a minimum publication threshold, or report confidence intervals following the bootstrapping methodology of Nakhoda, Whigham, and Zwanenburg (2023).
6. *Gaming via strategic co-authorship:* Adding co-authors from distant fields can inflate Δ without genuine integration. Mitigation: restrict Δ computation to corresponding-author publications; cross-check against S, which will remain low if the co-authored publications are incoherent.

The panel approach does not claim universal applicability. Single-publication assessments, highly collaborative fields where primary-category assignment is ambiguous, and emerging fields whose disciplinary boundaries are not yet reflected in existing similarity matrices all represent situations where the quantitative panel should be supplemented or replaced by expert judgment. Rafols (2019) has argued that indicators should be contextualized and subject to stakeholder validation; the framework presented here is designed in that spirit, providing structured quantitative input to — not a substitute for — informed evaluation.

Measurement in Practice

The taxonomy and panel presented above have implications for how interdisciplinarity is assessed in practice. This section provides operational guidance for implementing the indicator panel, drawing on the systematic five-step pipeline proposed by Cantone (2024): (1) unit of analysis selection, (2) taxonomy choice, (3) classification method, (4) operational definition, and (5) aggregation strategy. We organize the discussion around five procedural stages that a practitioner must navigate: data extraction, category mapping, similarity matrix construction, quality control, and software implementation.

Institutional Self-Assessment

Universities with full internal data access — including project records, publication databases, staff composition, and research budgets — are well-positioned to compute diversity and coherence indicators directly. The Rao-Stirling diversity index requires only a disciplinary classification of cited references and a similarity matrix; the coherence indicator requires bibliographic coupling data at the publication level. Both are computable from standard institutional repository data. The cross-field effect, however, requires citation data that institutions must typically obtain from external databases (Web of Science, Scopus, or the open-access OpenAlex). This data gap is the main practical obstacle to fully internal panel deployment. Where citation data is unavailable, a two-component profile (Δ, S) still provides useful discrimination between integrators and polymaths. Internal deployment of such a panel could support strategic self-assessment without relying on external ranking systems, in the spirit of responsible metrics advocated by Rafols (2019).

National Evaluation Agencies

A distinct challenge arises when a national evaluation agency must assess a researcher whose official classification is "Interdisciplinary" — that is, a researcher who does not fit any single disciplinary panel. Standard evaluation procedures assign reviewers from a single discipline, creating a structural mismatch. The indicator panel can support fairer evaluation by providing objective evidence of the type and degree of boundary-crossing: a high-Δ, high-S, high-E profile warrants reviewers from multiple fields, while a high-Δ, low-S profile may indicate a polymath who can be assessed field by field. The composition of the evaluation committee should reflect the structure revealed by the panel.

A concrete example is Spain's national research evaluation system. In 2023, the Comisión Nacional Evaluadora de la Actividad Investigadora (CNEAI) created *Campo 0: Interdisciplinar y Multidisciplinar*, the first dedicated evaluation track for interdisciplinary researchers within the six-year productivity assessment (*sexenio*). Mandated by Article 11.7 of Ley Orgánica 2/2023, which requires positive valuation of "the results of multidisciplinary and interdisciplinary research" across all fields, Campo 0 operationalizes precisely the multi/inter distinction discussed

above. Its criteria define *interdisciplinary* contributions as those "designed or structured by applying perspectives, theories, or methods associated with different disciplines" — an input-oriented definition measuring the integration of diverse methods into research design (mappable to high reference diversity Δ combined with high coherence S). Separately, *multidisciplinary* trajectories are recognized when supported by "at least two contributions in different disciplinary fields" — an output-oriented definition measuring publication breadth across fields (mappable to high variety $N \geq 2$). Notably, *transdisciplinary* research is absent from the Campo 0 criteria in all three editions published to date (2023–2025), consistent with the bibliometric measurement gap: transdisciplinarity, which involves non-academic partners and transcends disciplinary epistemologies, lacks standard bibliometric indicators. The panel (Δ, S, E) could operationalize the input-side measurement that ANECA's interdisciplinary track requires; publication field diversity would complement it for the output-side multidisciplinary criterion.

Data Extraction Procedures

The first practical step is to extract structured bibliographic records from a citation database. Three major sources are in current use. Web of Science (WoS) has been the standard choice for interdisciplinarity studies, offering Journal Subject Categories (approximately 254 categories in recent editions) organized into the Science Citation Index and Social Sciences Citation Index. Wang and Schneider (2020) used the combined JCR 2016 dataset, which covers 11,487 journals. Scopus provides an alternative with different category structure and broader coverage in some fields, while the open-access OpenAlex platform offers a community-maintained taxonomy with no subscription barrier.

The choice of database determines both the available metadata and the taxonomic structure. Key metadata fields include document type (articles are preferred; reviews and editorials are typically excluded unless specifically relevant), publication year (single-year snapshots or defined time windows), and — most critically — reference lists. References represent the "knowledge base" upon which a work is built and are preferred over citation counts for measuring knowledge integration, because they reflect deliberate intellectual choices by the authors rather than the post-publication reception of the work. Citation links, by contrast, are more appropriate for diffusion and impact measures but are dynamic and less stable over time.

For semantic or collaboration-based classification approaches, additional fields become relevant: title, abstract, and keywords support topic modeling and AI-based classification (Cantone, 2024), while author affiliations and co-authorship data enable organizational approaches. The volume of data can be substantial — the JCR 2016 dataset alone contains over 3 million inter-journal links and 50 million total citations (Leydesdorff, Wagner, and Bornmann, 2019), requiring attention to computational efficiency from the outset.

Category Mapping Protocols

Once bibliographic records are extracted, each publication must be assigned to one or more disciplinary categories. Four principal approaches exist, each with distinct trade-offs.

Journal-based assignment is the most common method. Each paper inherits the subject categories of its publishing journal. For multi-assigned journals, counts are either split proportionally or each category receives a full count. The proportion p_i of category i in a researcher's reference profile is then $p_i(x) = c_i(x)/\sum_j c_j(x)$, where $c_i(x)$ counts references in category i. This approach is stable, explainable, and computationally inexpensive, but it conflates journal disciplinarity with paper disciplinarity — an interdisciplinary paper published in a disciplinary journal inherits that journal's narrow classification.

Reference-based (cognitive) classification maps each cited reference to its journal's subject categories, building a disciplinary profile from the reference list rather than from the publishing journal. This is the most common approach for measuring knowledge integration (Cantone, 2024). A second-order variant uses references of references (Rafols and Meyer, 2009), but this does not resolve the fundamental asymmetry whereby the focal paper is treated as potentially interdisciplinary while its references are treated as mono-disciplinary by virtue of their journal assignments.

Semantic classification uses title, abstract, and keywords as input to supervised learning algorithms, topic models, or large language models. Cantone (2025) evaluated three LLMs for disciplinary classification and found that Gemini 1.5 Pro most closely approximated traditional citation-based assignments, ChatGPT 4o was most resilient to naming variations, and Claude 3.5 Sonnet offered a balanced profile. The advantages of semantic classification — intuitiveness and applicability to papers without clear journal assignments — are offset by limited explainability and sensitivity to prompt design.

Collaboration-based classification derives paper disciplinarity from the disciplinary identities of its authors, using degree background, departmental affiliation, or career trajectory. This approach faces severe recursive challenges (classifying an author requires classifying their prior work), low signal with small author counts, and ethical concerns about reducing individuals to disciplinary labels. It is most useful in institutional analyses where author-level metadata is available and well-curated.

The choice among these approaches is not neutral. Cantone (2024) observes that measures cluster by classification method rather than by conceptual dimension: journal-based and reference-based measures correlate with each other more strongly than either correlates with semantic-based measures of the same dimension. Practitioners should select the approach that best matches their research question and document the choice explicitly.

Similarity Matrix Construction

Diversity measures that incorporate disparity — including Rao-Stirling diversity and the DIV indicator — require a matrix of pairwise dissimilarity values d_{ij} between disciplinary categories. Not all cross-category links represent equal degrees of boundary-crossing: a reference from physics to mathematics represents less disparity than a reference from physics to art history. The construction of this matrix is therefore a critical methodological choice.

The standard approach computes cosine similarity from inter-category citation vectors. Wang and Schneider (2020) distinguished two variants. The Salton vector cosine $\mathrm{SC}(i,j) = \sum_k c_{ik}c_{jk} / \sqrt{\sum_k c_{ik}^2 \cdot \sum_k c_{jk}^2}$ uses the full citation profile of each category, where c_{ik} represents citations from category i to category k. The Ochiai binary cosine $\mathrm{SO}(i,j)$ uses a symmetrized version of direct cross-citation counts. A critical finding is that the correlation between the resulting dissimilarity measures $1 - \mathrm{SC}$ and $1 - \mathrm{SO}$ is only 0.54, and drops to 0.30 when using the inverse transformation $1/\mathrm{SC}$ versus $1/\mathrm{SO}$. The choice of similarity formula therefore dramatically affects diversity estimates.

Cosine similarity has several desirable properties for this application: it is non-parametric, bounded on $[0, 1]$, invariant to absolute scale (linear combinations preserve cosine values), and naturally disregards zero entries in sparse vectors (Leydesdorff, Wagner, and Bornmann, 2019). The standard procedure is to construct an $I \times I$ citation matrix between categories and convert similarities to dissimilarities via $d_{ij} = 1 - \cos(i,j)$.

When citation data is unavailable, alternative approaches exist. Cantone (2024) describes a confusion-matrix normalization suitable for classification schemes where misclassification probabilities are known. More recently, Cantone (2025) has explored large language model estimation, in which an LLM is prompted to provide similarity scores for all category pairs. This approach eliminates the need for citation database access but introduces new concerns: precision (variance across repeated identical queries), cross-model agreement (still limited), and robustness to trivial naming variations in category labels.

A further distributional concern affects the interpretation of results. Wang and Schneider (2020) found that Ochiai-based dissimilarity values ($1 - \mathrm{SO}$) are extremely left-skewed, with most values concentrated between 0.95 and 1.0. Under this distribution, Rao-Stirling diversity effectively reduces to the Simpson index because all cross-category pairs receive nearly identical disparity weights. Salton-based dissimilarity ($1 - \mathrm{SC}$) produces a more even distribution and thus preserves the intended role of the disparity component. Practitioners should examine the distribution of their chosen dissimilarity measure and avoid formulations that produce degenerate weighting.

Quality Control Procedures

The multiplicity of valid methodological choices creates a "researcher degrees of freedom" problem (Wang and Schneider, 2020): many defensible combinations of database, taxonomy, classification method, similarity formula, and aggregation level can yield substantially different results. A rigorous quality control protocol should address at least four levels.

Data integrity. Before computing any indicator, the analyst should validate journal-to-category assignments (checking multi-assignment logic), assess reference completeness (missing references bias diversity downward), define the citation window (retrospective all-time versus a fixed window such as five years), and document the treatment of self-citations.

Methodological consistency. Wang and Schneider (2020) demonstrated that measures which purport to capture the same dimension often exhibit surprisingly low correlations. Among their 23 indicator variants, measures clustered by methodological approach (overlap-based versus dissimilarity-based) rather than by conceptual dimension. Indicators that "should" correlate highly (for example, different operationalizations of diversity) sometimes showed correlations below 0.3. As a minimum validation step, any new indicator should be compared against established alternatives on the same dataset, with expected and actual correlations reported.

Aggregation level. A critical distinction separates elementary (paper-level) and collective (portfolio-level) measurement. In the elementary approach, each paper receives its own indicator score and the researcher's score is the mean or median. In the collective approach, all references from all papers are pooled into a single distribution before computing the indicator. Wang and Schneider (2020) found that elementary and collective Rao-Stirling diversity values correlate at 0.91 when the same dissimilarity measure is used, but correlation drops to 0.18 when different dissimilarity formulas are applied at the same aggregation level — confirming that the similarity matrix, not the aggregation level, is the dominant source of variation. The choice of level should be guided by the research question: elementary measurement captures the typical paper, while collective measurement characterizes the overall knowledge base.

Granularity sensitivity. Finer taxonomies (such as 254 WoS subject categories) yield higher measured interdisciplinarity than coarser taxonomies (such as 40 OECD categories), because more category boundaries are available to cross (Cantone, 2024). There is no universally "correct" granularity; the appropriate level depends on the purpose of the analysis. As a robustness check, practitioners should repeat the analysis at multiple granularity levels and report whether conclusions are stable.

An additional form of validation compares diversity measures against independent structural indicators. Leydesdorff, Wagner, and Bornmann (2019) propose comparing Rao-Stirling diversity with betweenness centrality in the journal

citation network. High correlation would provide convergent evidence that both capture aspects of interdisciplinary positioning; divergence would indicate that the measures are capturing different phenomena. Comparing citing-side (knowledge integration) and cited-side (knowledge diffusion) indicators separately can further clarify which dimension is being measured.

Uncertainty quantification. Even when methodological choices are held fixed, the stochastic nature of reference lists introduces measurement uncertainty. Nakhoda, Whigham, and Zwanenburg (2023) proposed a non-parametric bootstrap approach to quantify this uncertainty for the Rao-Stirling index. Their procedure takes a publication's N recognized subject-category assignments, resamples them with replacement to produce $B = 500$ bootstrap replicates of size N, computes the Rao-Stirling index for each resample, and constructs a bias-corrected 95% confidence interval from the resulting distribution. Across 42,660 publications, the median confidence-interval width was approximately 0.15, but values ranged from zero (when all references fell in a single category) to over 0.6. Papers with fewer than ten categorized references exhibited particularly wide intervals, indicating that point estimates of interdisciplinarity are unreliable for short reference lists. The authors further showed that combining the bootstrap confidence interval with the number of references yields a more effective reliability filter than either criterion alone, enabling practitioners to flag publications whose interdisciplinarity scores should not be interpreted at face value.

Software Implementation

Several software resources support the computation of interdisciplinarity indicators. Leydesdorff, Wagner, and Bornmann (2019) provide a publicly available routine that accepts citation matrices in Pajek format and computes Rao-Stirling diversity, the DIV indicator, Gini coefficient, Simpson index, Shannon entropy, and separate disparity and variety components. Wang and Schneider (2020) combined SQL queries on an in-house WoS database with the R package *sna* for betweenness centrality and custom R scripts for other measures. At present, no single integrated package covers the full indicator panel proposed in this review; assembling one from existing components is a natural next step.

Computational scalability requires attention when working with large datasets. The similarity matrix involves $O(n^2)$ pairwise comparisons for n categories — manageable for 254 WoS categories (approximately 32,000 pairs) but potentially expensive if finer-grained taxonomies are used. Sparse matrix representations are appropriate because most inter-category citation counts are zero. The similarity matrix should be precomputed once and reused across all papers. Paper-level indicator calculations are independent and thus naturally parallelizable.

Reproducibility demands that every methodological choice be documented: database version (e.g., JCR 2016 or JCR 2023), taxonomy and its granularity, classification method, similarity formula and dissimilarity transformation, aggregation level, and any filtering criteria applied. Sensitivity analyses — repeating

the computation with alternative choices at each stage — should accompany the main results (Cantone, 2024). A practical implementation strategy is to begin with a single year, a single taxonomy, and reference-based classification, validate against published results on comparable data, and only then introduce additional complexity. Modular pipeline design, with separate extraction, classification, and measurement stages, facilitates both incremental validation and the eventual substitution of components as methods improve.

Open Problems and Future Directions

Several important questions remain unresolved and merit further investigation.

First, the relationship between self-reported and bibliometric interdisciplinarity is poorly understood. Aksnes, Karlstrøm, and Piro (2026), surveying over 3,000 publications across all fields, found that self-reported and bibliometric interdisciplinarity measures "rarely correspond." Testing Shannon entropy, the true diversity measure (^2D_S), and the DIV* decomposition against researcher self-assessments, they obtained correlations ranging from 0.13 to 0.18, explaining only 2–3% of variance. Researchers assess interdisciplinarity based on collaboration dynamics and methodological integration, not reference patterns. This raises fundamental questions about construct validity: if even a battery of complementary bibliometric indicators fails to capture what researchers themselves mean by interdisciplinarity, the gap must be acknowledged in any evaluation framework.

Second, the estimation of disciplinary similarity matrices — a critical input to Rao-Stirling diversity and related measures — has traditionally relied on citation coupling data. Cantone (2025) has recently explored the use of large language models (ChatGPT 4o, Claude 3.5 Sonnet, Gemini 1.5 Pro) to estimate similarity matrices directly from disciplinary labels, finding partial agreement with citation-based estimates. If validated, this approach could reduce the data requirements for computing diversity indicators, though robustness to trivial naming variations remains a concern.

Third, uncertainty quantification for interdisciplinarity measures is largely absent from the literature. Point estimates of diversity or coherence are reported without confidence intervals, making it difficult to assess whether observed differences between researchers or institutions are statistically meaningful. Nakhoda, Whigham, and Zwanenburg (2023) identified three sources of uncertainty in citation-based measures — arbitrary referencing behavior, uncategorized references, and invalid journal-to-paper category inheritance — and proposed a bootstrapping method to estimate confidence intervals for the Rao-Stirling index. Their finding that confidence intervals can span up to 0.6 points underscores the risk of over-interpreting small differences in diversity scores.

Fourth, the relationship between interdisciplinarity and research quality remains contested. Citation-based quality measures appear to penalize interdisciplinary

work in the short term: balanced disciplinary portfolios are associated with lower citation counts over typical evaluation windows (Cantone, 2024, citing Larivière and Gingras, 2010). However, recent evidence suggests that IDR achieves greater and more lasting impact over longer time horizons — the "penalty" is better understood as a diffusion lag across disciplinary communities. The type of interdisciplinarity also matters: Xiang, Romero, and Teplitskiy (2025), analyzing 128,950 manuscripts across 62 journals, found that knowledge-base interdisciplinarity (diverse references) is associated with higher acceptance rates, while topic interdisciplinarity (crossing disciplinary subject boundaries) is associated with lower acceptance rates — the two dimensions have opposite effects on peer review outcomes. A panel that characterizes the *type* of boundary-crossing — as ours does — provides the structural context needed to interpret quality indicators correctly. Importantly, Xiang et al. also found that journals designated as "interdisciplinary" by their publisher showed no penalty against either form of interdisciplinarity, suggesting that the observed biases are specific to disciplinary venues rather than inherent to interdisciplinary work itself. This finding reinforces the case for dedicated interdisciplinary evaluation contexts. Notably, our cross-field effect E is defined as a fraction (not an absolute citation count), avoiding the conflation of citation volume with interdisciplinarity that affects some diffusion measures.

Fifth, the distribution-based approaches of Zhou et al. (2023) offer a promising direction for enriching scalar panels. Their IKF framework decomposes what Rao-Stirling aggregates, revealing *which* disciplines contribute to measured diversity and *how deeply* each contributes. Integrating such distributional information into practical evaluation frameworks is an open challenge.

Sixth, participatory and design-informed approaches offer a methodological alternative to purely quantitative indicator deployment. Marres and de Rijcke (2020) propose "engaging indicators" that recognize indicators' dual role: not only representing research patterns but also organizing communities of interpretation. Their methodology combines scientometric analysis with stakeholder workshops and interactive mapping, emphasizing that indicators are designed entities whose material and interactive forms include or exclude actors in evaluation processes. This approach addresses Rafols' (2019) concern that indicators, originally developed as tools to inform decision-making, risk becoming "ignorance-producing devices" when deployed mechanically without contextual interpretation. The challenge lies in scaling participatory methods — which are labor-intensive and context-specific — to institutional and national evaluation frameworks while preserving their capacity to surface contested meanings of interdisciplinarity.

Seventh, the measurement of *transdisciplinary* research remains an open frontier. The OECD tripartite typology and its elaboration by Wagner et al. (2011) distinguish transdisciplinarity from multi- and interdisciplinarity by its integration of disciplinary epistemologies and, increasingly, by its engagement with non-academic partners (Borlaug and Svartefoss, 2025). Yet no standard bibliometric indicator captures this dimension. Diversity indicators measure breadth

of knowledge inputs; coherence indicators measure integration of the knowledge base; but neither detects whether research transcends academic boundaries to engage practitioner knowledge, policy contexts, or community stakeholders. This measurement gap has practical consequences: Spain's ANECA, in designing Campo 0 for interdisciplinary evaluation, explicitly covers "interdisciplinar y multidisciplinar" research but omits transdisciplinary criteria entirely (BOE-A-2023-25537 through BOE-A-2025-26118) — a pragmatic acknowledgment that what cannot be measured should not be required. Future work might explore hybrid approaches combining bibliometric panels with qualitative evidence of stakeholder engagement, along the lines of Marres and de Rijcke's (2020) participatory indicators, to bridge this gap.

Case Study: Department-Level Evaluation

The toy-data demonstration in Section 4 established that the three-component panel can distinguish researcher archetypes in principle. We now apply the panel to a more realistic scenario: evaluating the interdisciplinarity of seven researchers in a university physics and materials science department. This case study illustrates the panel's practical operation at a scale that mirrors real institutional assessment exercises, using a richer set of disciplinary categories, realistic publication volumes, and supplementary institutional data.

Department Context and Data

Consider a hypothetical department of Physics and Materials Science at a mid-sized research university. The department houses seven researchers at various career stages: three senior faculty (twelve to fifteen years post-PhD), two mid-career faculty (eight to nine years), and two early-career researchers (four to five years). The evaluation question is whether each researcher qualifies for an interdisciplinary research funding track, a classification that carries consequences for reviewer assignment, panel composition, and reporting expectations.

The disciplinary landscape is captured by six Web of Science subject categories relevant to the department's research portfolio:

Category	Label	Relation to department
C_1	Physics, condensed matter	Core
C_2	Materials science	Core
C_3	Chemistry, physical	Adjacent
C_4	Optics	Adjacent
C_5	Engineering, electrical	Adjacent
C_6	Nanoscience and nanotechnology	Adjacent

The pairwise similarities between these categories, derived from cosine similarity on the WoS journal citation network, are:

	C_1	C_2	C_3	C_4	C_5	C_6
C_1	1.00	0.60	0.40	0.35	0.30	0.55
C_2	0.60	1.00	0.50	0.25	0.40	0.65
C_3	0.40	0.50	1.00	0.30	0.20	0.45
C_4	0.35	0.25	0.30	1.00	0.45	0.35
C_5	0.30	0.40	0.20	0.45	1.00	0.50
C_6	0.55	0.65	0.45	0.35	0.50	1.00

This matrix exhibits a richer structure than the toy example of Section 4: the core categories (C_1, C_2) have high mutual similarity (0.60) and moderate connections to adjacent fields; nanoscience (C_6) is intrinsically interdisciplinary, with above-average similarity to five of the six categories; and the most distant pair is chemistry–engineering ($s_{35} = 0.20$), reflecting genuine epistemological distance.

Each researcher's publication portfolio is summarized by a reference distribution vector $\mathbf{p}$ over the six categories, derived from the aggregate reference lists of all publications in the assessment period. Table 1 presents these profiles alongside publication counts, citation totals, and the fraction of citations received from outside each researcher's primary category.

Table 1. Researcher profiles. Career stage abbreviations: S = senior, M = mid-career, E = early-career. The reference vector $\mathbf{p} = (p_1, \ldots, p_6)$ gives the proportion of references to each category.

Researcher	Stage	Pubs	$\mathbf{p}$	Cites	Cross-field cites
Chen	S	42	(0.35, 0.30, 0.20, 0.05, 0.05, 0.05)	520	180 (35%)
Al-Rahman	M	35	(0.20, 0.20, 0.20, 0.20, 0.15, 0.05)	280	30 (11%)
Kowalski	E	18	(0.65, 0.25, 0.06, 0.02, 0.01, 0.01)	85	10 (12%)
Nguyen	S	55	(0.25, 0.25, 0.15, 0.10, 0.15, 0.10)	890	420 (47%)
Romero	M	28	(0.10, 0.15, 0.05, 0.05, 0.10, 0.55)	195	140 (72%)
Karlsson	E	12	(0.50, 0.30, 0.10, 0.05, 0.03, 0.02)	45	8 (18%)
Osei	S	48	(0.70, 0.20, 0.05, 0.03, 0.01, 0.01)	680	75 (11%)

Several features of this dataset merit comment. Al-Rahman's reference distribution is nearly uniform across the first four categories, resembling the polymathic archetype of Section 4. Romero's distribution is dominated by a single adjacent category (C_6, nanoscience), yet her cross-field citation fraction is the highest in the department. These contrasting profiles foreshadow the discriminations that the panel will reveal.

Panel Computation

We trace the computation of all three panel components for the full set of researchers.

Diversity (Δ). Applying the Rao-Stirling formula $\Delta = \sum_{i,j}(1 - s_{ij})\, p_i\, p_j$ to each reference vector yields the values in Table 2. We illustrate the computation for two researchers whose profiles are of particular evaluative interest.

For Nguyen ($\mathbf{p}_N = (0.25, 0.25, 0.15, 0.10, 0.15, 0.10)$), the off-diagonal sum involves fifteen distinct pairs. The dominant contributions come from the high-weight, high-distance pairs: the (C_1, C_5) term contributes $2 \times 0.25 \times 0.15 \times 0.70 = 0.053$; the (C_1, C_4) term contributes $2 \times 0.25 \times 0.10 \times 0.65 = 0.033$; and the (C_3, C_5) term contributes $2 \times 0.15 \times 0.15 \times 0.80 = 0.036$. Summing all fifteen pairs gives $\Delta_N = 0.464$.

For Al-Rahman ($\mathbf{p}_F = (0.20, 0.20, 0.20, 0.20, 0.15, 0.05)$), the near-uniform distribution generates many terms of comparable magnitude. The total evaluates to $\Delta_F = 0.610$, the highest in the department — a consequence of dispersed weight across categories with substantial mutual distances.

Coherence (S). The mean bibliographic coupling strength is computed from pairwise cosine similarities between publication reference vectors. This indicator captures whether a researcher's diverse publications form an integrated whole or represent disconnected contributions.

Cross-field effect (E). The fraction of citations received from outside each publication's primary category is pooled across the researcher's entire portfolio.

Table 2. Panel values for all seven researchers.

Researcher	Δ	S	E	Pattern
Chen	0.390	0.42	0.35	Moderate integrator
Al-Rahman	0.610	0.04	0.11	Polymath
Kowalski	0.190	0.53	0.12	Specialist
Nguyen	0.464	0.50	0.47	Strong integrator
Romero	0.374	0.50	0.72	Niche-bridge specialist
Karlsson	0.244	0.58	0.18	Early-career specialist
Osei	0.176	0.58	0.11	Specialist

The range of diversity values (0.176 to 0.610) is wider than in the toy example, reflecting both genuine disciplinary variation and the effect of a larger category set with heterogeneous pairwise distances. Coherence values separate cleanly into two groups: Al-Rahman's near-zero coherence ($S = 0.04$) stands in sharp contrast to the moderate-to-high coherence of all other researchers ($S \geq 0.42$), indicating that his diverse publications share essentially no references.

Interpretation and Classification

The panel values in Table 2 support a structured classification of each researcher against the evaluation question. We adopt indicative decision thresholds: $\Delta \geq 0.40$ for substantial diversity, $S \geq 0.30$ for meaningful coherence, and $E \geq 0.30$

for significant cross-field impact. These thresholds are illustrative; in practice, they would be calibrated to the local disciplinary context (see Section 5).

Nguyen ($\Delta = 0.464$, $S = 0.50$, $E = 0.47$): genuine integrator. Nguyen exceeds all three thresholds comfortably. His publications span four major categories with substantial weight, yet they share a common reference base ($S = 0.50$) that indicates systematic knowledge integration rather than disconnected forays. Nearly half of his citations (47%) originate outside the primary category, confirming that his work achieves genuine cross-field impact. Institutional data reinforce this assessment: 22 of 55 publications involve co-authors from other departments (chemistry, engineering, medical school), and all four of his grants were awarded through interdisciplinary funding panels. The panel recommends classifying Nguyen for the interdisciplinary track, with a review committee spanning at least three of his active categories.

Chen ($\Delta = 0.390$, $S = 0.42$, $E = 0.35$): borderline integrator. Chen falls just below the diversity threshold ($\Delta = 0.390$ versus the 0.40 cut-off) but meets the coherence and cross-field effect criteria. Her publication pattern — strong in condensed matter and materials science, with systematic engagement in physical chemistry — suggests an emerging integrator whose interdisciplinary reach is concentrated among closely related fields. The moderate disparity among her active categories (most pairwise similarities exceed 0.35) keeps her Rao-Stirling value below the threshold, even though her research practice is substantively cross-disciplinary. The panel recommends classification as an emerging integrator, with periodic reassessment.

Romero ($\Delta = 0.374$, $S = 0.50$, $E = 0.72$): niche-bridge specialist. Romero presents the most instructive case for panel interpretation. Her diversity is below the threshold, reflecting a concentrated position in nanoscience ($p_6 = 0.55$). Yet her cross-field effect is exceptional: 72% of her citations come from outside nanoscience, indicating that her specialized work serves as a bridge connecting nanoscience to condensed matter, materials science, and engineering. Her coherence ($S = 0.50$) confirms that this bridge role is sustained through an integrated research program, not occasional cross-field publications. Co-authorship data corroborate the interpretation: 18 of 28 publications involve collaborators from other departments. The panel correctly identifies Romero as a case requiring expert review — she does not fit the standard integrator profile, but her structural role in the departmental research network may be equally valuable for interdisciplinary funding purposes.

Al-Rahman ($\Delta = 0.610$, $S = 0.04$, $E = 0.11$): polymath. Al-Rahman has the highest diversity in the department, with near-uniform weight across four categories. Yet his coherence is essentially zero ($S = 0.04$), indicating that his publications in different fields share no common reference base — each constitutes an independent contribution to a separate disciplinary conversation. His cross-field effect is correspondingly low (11%): despite publishing broadly, his work is cited almost exclusively within the category where each paper was published. This is the classic breadth-without-integration pattern identified by our toy

example. Notably, Al-Rahman has zero cross-departmental co-authorships, confirming that his disciplinary breadth does not translate into collaborative integration. The panel recommends standard disciplinary evaluation, not the interdisciplinary track.

Kowalski, Karlsson, and Osei: specialists. The remaining three researchers have diversity values below 0.35, placing them clearly in the specialist category. Osei ($\Delta = 0.176$) is the most concentrated, with 70% of references in condensed matter physics; his high coherence ($S = 0.58$) and low cross-field effect ($E = 0.11$) describe a focused and productive disciplinary researcher. Kowalski and Karlsson are early-career researchers whose low diversity reflects limited publication volume rather than a settled disciplinary profile. For Karlsson (12 publications), bootstrap confidence intervals yield $\Delta \in [0.18, 0.31]$ at the 95% level, suggesting that quantitative panel assessment should be deferred until his portfolio reaches approximately 20 publications. All three are appropriately routed to standard disciplinary evaluation.

A noteworthy outcome is the institutional data on co-authorship diversity. For all seven researchers, the fraction of cross-departmental publications aligns with the panel classification: Nguyen and Chen have substantial cross-departmental collaboration (40% and 55% respectively), Romero has the highest rate (64%), while Al-Rahman, Osei, Kowalski, and Karlsson have zero cross-departmental papers. This convergence between bibliometric indicators and institutional process data increases confidence in the panel's classifications.

Comparison with Single-Indicator Approaches

The case study provides a concrete demonstration of why single-indicator approaches are inadequate. We compare the panel classification against three single-indicator rankings.

Shannon entropy ($H = -\sum p_i \log_2 p_i$) applied to the reference distributions yields the following ranking: Al-Rahman ($H = 2.43$), Nguyen (2.32), Chen (2.05), Romero (1.85), Karlsson (1.45), Kowalski (1.22), Osei (1.08). Under this measure, Al-Rahman — the polymath with zero integration — ranks as the most interdisciplinary researcher in the department, ahead of Nguyen, the genuine integrator.

Rao-Stirling diversity alone produces a qualitatively similar ranking: Al-Rahman ($\Delta = 0.610$), Nguyen (0.464), Chen (0.390), Romero (0.374), Karlsson (0.244), Kowalski (0.190), Osei (0.176). Again, Al-Rahman leads. Both diversity-only approaches reward breadth irrespective of whether that breadth is accompanied by knowledge integration.

Cross-field citation ratio alone (E) reorders the ranking substantially: Romero ($E = 0.72$), Nguyen (0.47), Chen (0.35), Karlsson (0.18), Kowalski (0.12), Al-Rahman (0.11), Osei (0.11). This measure correctly demotes Al-Rahman but elevates Romero — a specialist in nanoscience — to the top

position, conflating bridge-specialist impact with genuine integrative research.

Each single indicator produces a different "most interdisciplinary" researcher, and each misclassifies at least one profile. The full panel avoids these errors because it operates on three orthogonal dimensions simultaneously. Al-Rahman's polymath profile (Δ high, S near zero, E low) is unambiguously detected; Romero's bridge-specialist role (Δ moderate, S moderate, E very high) is flagged for expert review rather than automatic classification; and Nguyen's integrator status (Δ, S, and E all above threshold) is confirmed with high confidence. This three-way discrimination is the panel's primary practical advantage.

Limitations

Several limitations of the case study should be acknowledged, as they illustrate broader challenges for panel deployment.

Threshold calibration. The decision thresholds used here ($\Delta \geq 0.40$, $S \geq 0.30$, $E \geq 0.30$) are illustrative, not empirically validated. Their calibration requires benchmarking against cases with known interdisciplinary status — for instance, researchers funded through interdisciplinary mechanisms whose work has been independently evaluated by expert panels. Until such benchmarking is undertaken, the thresholds should be treated as adjustable parameters that institutions set according to local disciplinary norms.

Similarity matrix temporality. The similarity matrix is derived from the WoS 2016 journal citation network, while the assessment covers publications through 2025. Disciplinary boundaries shift over time: nanoscience (C_6), for instance, may have been more distinct from condensed matter physics (C_1) a decade ago than it is today. Using a static similarity matrix introduces a systematic bias that particularly affects researchers working at the boundaries of rapidly converging fields. An updated matrix, computed from citation data contemporaneous with the assessment period, would mitigate this concern.

Early-career instability. For Karlsson (12 publications) and Kowalski (18 publications), the panel values are computed from relatively sparse data. Bootstrap resampling for Karlsson yields 95% confidence intervals of $\Delta \in [0.18, 0.31]$, a range that spans the boundary between specialist and moderate diversity. More generally, the coherence indicator S is sensitive to portfolio size because the number of pairwise comparisons grows quadratically with the number of publications. For small portfolios, a single atypical publication can substantially alter S. A practical recommendation is to supplement panel scores with confidence intervals and to defer classification decisions for portfolios below approximately 20 publications.

Temporal aggregation. The case study aggregates each researcher's full career output, masking potentially important trajectories. Kowalski's early publications are concentrated in condensed matter, but her most recent work shows expansion into materials science and physical chemistry — a trajectory that career-level

aggregation obscures. A windowed variant of the panel (e.g., computed over a rolling three-year window) would capture such dynamics, at the cost of reduced statistical stability for researchers with lower annual publication rates.

Bibliographic coupling limitations. The coherence indicator S measures integration through shared references. Al-Rahman's near-zero coherence ($S = 0.04$) may understate latent methodological connections between his publications if those connections operate through shared techniques or concepts rather than shared literature. Text-based measures — such as co-word analysis of abstracts or topic model similarity — could complement bibliographic coupling in cases where methodological integration is suspected but not reflected in reference overlap.

Conclusions

The bibliometric measurement of interdisciplinary research remains an unsettled problem. Our review of the indicator landscape reveals a field heavily concentrated on diversity measures — particularly Rao-Stirling and its variants — while coherence, diffusion, and novelty dimensions receive comparatively little attention. The empirical evidence, especially Wang and Schneider's (2020) finding of low consistency across 23 measures and Leydesdorff et al.'s (2019) demonstration of limited discriminatory power, strongly suggests that no single indicator is adequate.

This framing aligns with the four motivating questions stated for this project. The panel is designed to separate impact/quality evidence from simple breadth claims (OQ1), to remain computable with institutional data plus clearly stated external dependencies (OQ2), to distinguish cross-disciplinary integration from polymathic accumulation (OQ3), and to support auditable agency-level evaluation protocols through explicit multidimensional evidence rather than single-score ranking (OQ4).

The three-component panel we propose — diversity (Δ), coherence (S), and cross-field effect (E) — addresses this inadequacy by spanning three orthogonal dimensions. Our toy-data demonstration shows that the panel uniquely characterizes integrators, polymaths, and specialists where any single component fails. The department-level case study of Section 7 confirms this discriminatory power at realistic scale: the panel correctly identifies Al-Rahman as a polymath despite his having the highest diversity score, flags Romero's bridge-specialist role for expert review, and confirms Nguyen's integrator status across all three dimensions — classifications that no single indicator achieves. The analytical robustness result — that the diversity-based discrimination is preserved under perturbations of up to 35% of the similarity matrix — provides confidence that the approach is not an artifact of parameter tuning.

Rafols (2019) has argued persuasively that science and technology indicators should be contextualized, multidimensional, and subject to stakeholder validation.

Our panel is designed in this spirit: it presents three separate dimensions rather than collapsing them into a single score, and its interpretation depends on the evaluation context. The practical deployment of such panels — whether for institutional self-assessment or national agency review — requires attention to the methodological choices surveyed in Section 5, the open problems identified in Section 6, and the practical lessons illustrated in Section 7. Empirical validation on real university data, building on the illustrative case study presented here, is the natural next step.

References

- Abramo, G., D'Angelo, C. A., and Zhang, L. (2018). A comparison of two approaches for measuring interdisciplinary research output: The disciplinary diversity of authors vs the disciplinary diversity of the reference list. *Journal of Informetrics*, 12(4):1182–1193.
- Aksnes, D. W., Karlstrøm, H., and Piro, F. N. (2026). Self-reported and bibliometric interdisciplinarity measures rarely correspond: a survey-based comparative analysis of indicators and researcher perceptions. *Scientometrics*, 131:189–208.
- Bollen, J., Van de Sompel, H., Hagberg, A., and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE*, 4(6):e6022.
- Borlaug, S. B. and Svartefoss, S. M. (2025). Evaluating transdisciplinary research quality. In Sivertsen, G. and Langfeldt, L. (eds.), *Challenges in Research Policy*, pp. 13–20. Springer, Cham.
- Bornmann, L., Tekles, A., Zhang, H. H., and Ye, F. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics*, 13(4):100979.
- Boyack, K. W., Klavans, R., and Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3):351–374.
- Cantone, G. G. (2024). How to measure interdisciplinary research? A systemic design for the model of measurement. *Scientometrics*, 129:4937–4982.
- Cantone, G. G. (2025). Estimation of disciplinary similarity with large language models. *Scientometrics*, 130(10):5345–5373.
- Choi, B. C. K. and Pak, A. W. P. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clinical and Investigative Medicine*, 29(6):351–364.
- Fontana, M., Iori, M., Montobbio, F., and Sinatra, R. (2020). New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy*, 49(7):104063.
- Goyanes, M., Demeter, M., Grané, A., Albarrán-Lozano, I., and Gil de Zúñiga, H. (2020). A mathematical approach to assess research diversity:

Operationalization and applicability in communication sciences, political science, and beyond. *Scientometrics*, 125(3):2299–2322.
- Hammarfelt, B. (2020). Discipline. *Knowledge Organization*, 47(3):244–256.
- Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2):427–432.
- Jensen, P. and Lutkouskaya, K. (2014). The many dimensions of laboratories' interdisciplinarity. *Scientometrics*, 98(1):619–631.
- Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2):363–375.
- Jost, L. (2009). Mismeasuring biological diversity: Response to Hoffmann and Hoffmann (2008). *Ecological Economics*, 68(4):925–928.
- Klein, J. T. (2008). Evaluation of interdisciplinary and transdisciplinary research: A literature review. *American Journal of Preventive Medicine*, 35(2S):S116–S123.
- Larivière, V. and Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology*, 61(1):126–131.
- Leinster, T. and Cobbold, C. A. (2012). Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489.
- Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal–journal citation relations using the Journal Citation Reports? *Journal of the American Society for Information Science and Technology*, 57(5):601–613.
- Leydesdorff, L. and Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, 5(1):87–100.
- Leydesdorff, L., Wagner, C. S., and Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*, 13(1):255–269.
- Marres, N. and de Rijcke, S. (2020). From indicators to indicating interdisciplinarity: A participatory mapping methodology for research communities in-the-making. *Quantitative Science Studies*, 1(3):1041–1055.
- Morillo, F., Bordons, M., and Gómez, I. (2001). An approach to interdisciplinarity through bibliometric indicators. *Scientometrics*, 51(1):203–222.
- Morillo, F., Bordons, M., and Gómez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology*, 54(13):1237–1249.
- Moulton, R. and Jiang, Y. (2018). Maximally consistent sampling and the Jaccard index of probability distributions. In *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*, pages 347–356. IEEE.
- Mutz, R. (2022). Diversity and interdisciplinarity: Should variety, balance and disparity be combined as a product or better as a sum? An information-theoretical and statistical estimation approach. *Scientometrics*,

127(12):7397–7414.

- Nakhoda, M., Whigham, P., and Zwanenburg, S. (2023). Quantifying and addressing uncertainty in the measurement of interdisciplinarity. *Scientometrics*, 128:6107–6127.
- Noji, H., Yasuda, R., Yoshida, M., and Kinosita, K. (1997). Direct observation of the rotation of F1-ATPase. *Nature*, 386(6622):299–302.
- OECD (1998). *Interdisciplinarity in Science and Technology*. Paris: OECD Directorate for Science, Technology and Industry.
- Porter, A. L. and Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81(3):719–745.
- Rafols, I. (2019). S&T indicators in the wild: contextualization and participation for responsible metrics. *Research Evaluation*, 28(1):7–22.
- Rafols, I. and Meyer, M. (2009). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2):263–287.
- Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15):707–719.
- Stokols, D., Fuqua, J., Gress, J., Harvey, R., Phillips, K., Baezconde-Garbanati, L., Unger, J., Palmer, P., Clark, M. A., Colby, S. M., Morgan, G., and Trochim, W. (2003). Evaluating transdisciplinary science. *Nicotine & Tobacco Research*, 5(Suppl 1):S21–S39.
- Tomishige, M., Klopfenstein, D. R., and Vale, R. D. (2002). Conversion of Unc104/KIF1A kinesin into a processive motor after dimerization. *Science*, 297(5590):2263–2267.
- Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157):468–472.
- Wang, J., Veugelers, R., and Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436.
- Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., and Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1):14–26.
- Wang, Q. and Schneider, J. W. (2020). Consistency and validity of interdisciplinarity measures. *Quantitative Science Studies*, 1(1):239–263.
- Xiang, S., Romero, D. M., and Teplitskiy, M. (2025). Evaluating interdisciplinary research: Disparate outcomes for topic and knowledge base. *Proceedings of the National Academy of Sciences*, 122(16):e2409752122.
- Zhang, L., Rousseau, R., and Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology*, 67(5):1257–1265.
- Zhou, Q., Guns, R., and Engels, T. C. E. (2023). Towards indicating interdisciplinarity: Characterizing interdisciplinary knowledge flow. *Journal of*

the Association for Information Science and Technology, 74(11):1325–1340.

- Zwanenburg, S., Nakhoda, M., and Whigham, P. (2022). Toward greater consistency and validity in measuring interdisciplinarity: a systematic and conceptual evaluation. *Scientometrics*, 127:3035–3065.