Visible Thinking in Conversation: Ecological Validity of the Expressive Cognition Verbal Reasoning Index in a High-Ability Naturalistic Sample

A research paper from expressivecognition.org

CONFLICT OF INTEREST STATEMENT

This research was conducted by the developer of the Expressive Cognition assessment tool (expressivecognition.org), a freely accessible assessment with an optional paid report tier. This relationship is disclosed in the interest of full transparency. No external funding was received for this research.

AI USAGE STATEMENT

This study employs Claude Sonnet 4 (Anthropic) as an automated scoring agent within a predefined behavioral rubric and a blinded multi-pass evaluation protocol. The model functions as a measurement instrument: it applies the rubric to speech transcripts and generates dimension-level scores and supporting evidence under controlled prompting conditions. Claude was also used for limited drafting and editorial assistance in the preparation of this manuscript. The theoretical framework, research design, analyses, and all interpretive conclusions are the work of the Expressive Cognition research program. Full responsibility for the accuracy, integrity, and originality of the manuscript rests with the project.

ABSTRACT

Construct validity studies typically test whether an instrument detects differences between groups *known* to differ. This study asks a harder question: within a naturalistic sample of high-ability speakers drawn from a single conversational context, does the Expressive Cognition (EC) Verbal Reasoning Index detect meaningful dimensional variation, and does that variation correlate with an external criterion? Thirty guests from *Conversations with Tyler*, a long-form interview podcast, were scored blinded across six core dimensions and two moderators using the EC behavioral descriptor rubric applied by Claude Sonnet 4 in three independent passes per guest. External intellectual reputation scores (1--10) were assigned before scoring. The VRI composite correlates weakly with reputation ($r = 0.298$), attenuated by ceiling effects in this pre-selected sample. However, two specific dimensions emerge as moderate-to-strong predictors: Generative Self-Monitoring ($r = 0.441$, $p = .015$) and Epistemic Calibration ($r = 0.420$, $p = .021$). These dimensions capture what this paper terms *generative intelligence* — the visible, real-time process of self-correction and epistemic marking in spontaneous speech. A supplementary written-vs-spoken pilot ($n = 3$) provides convergent validity evidence: dimensions that measure the person (Abstraction, Vocabulary) are stable across modalities, while dimensions that measure the reasoning process (Generative Self-Monitoring, Conceptual Continuity) shift in theoretically predicted directions. Cross-study comparison with prior construct validity work (companion paper) reveals systematic context-sensitivity: Generative Self-Monitoring is null in adversarial legal argument (SCOTUS studies), the second-strongest differentiator in academic seminar speech (MICASE study), and the strongest predictor of reputation in conversational interview (present study) — a pattern predicted by Levelt's (1983) self-monitoring model. The findings suggest

that what distinguishes high-reputation thinkers in spontaneous conversational speech is not what they know — Abstraction and Compression ceiling out — but how visibly they think.

Keywords: ecological validity, verbal reasoning, spontaneous speech, LLM scoring, intellectual reputation, self-monitoring, epistemic calibration, conversational discourse

INTRODUCTION

From Construct Validity to Ecological Validity

A companion paper established that the Expressive Cognition (EC) scoring rubric detects known differences in verbal reasoning quality in the predicted direction across two naturalistic speech corpora — Supreme Court oral arguments and university academic seminars. That paper's claim was deliberately narrow: known-groups validity, the most basic form of construct evidence. The present study extends the validity argument by testing whether EC scores correlate with an external criterion — intellectual reputation — in a naturalistic, uncontrolled conversational context where the instrument had no involvement in the speech elicitation.

This shift from known-groups to criterion-related validity imposes additional demands. Known-groups designs provide clear predicted rank orders: elite advocate should outscore first-time advocate; faculty should outscore undergraduate. In the present design, no such rank order is guaranteed. All 30 speakers are accomplished public intellectuals interviewed by the same host; the question is whether the scoring rubric detects *within-group* variation that tracks something external, or whether it compresses this pre-selected sample into indistinguishable scores.

The Conversational Context as a Validity Test

Conversations with Tyler (CWT) is a long-form interview podcast hosted by economist Tyler Cowen. Each episode features a single guest in a wide-ranging, intellectually demanding conversation. The host's interviewing style is distinctive: questions are rapid, unexpected, and often push guests outside their primary domain of expertise. This creates natural "pivot moments" — points in the conversation where the guest must reason in real time about something they have not prepared for.

This context provides a validity test that neither the SCOTUS nor MICASE corpora could. SCOTUS oral arguments are adversarial and rehearsed; MICASE seminars are exploratory but institutionally structured. CWT interviews are conversational, unrehearsed, and intellectually challenging — closer to the kind of spontaneous reasoning EC was designed to measure. If the rubric detects meaningful variation here, the construct generalizes to a third fundamentally different speech context.

The External Criterion: Intellectual Reputation

Each guest was assigned a reputation score (1--10) based on an assessment of citation impact, public intellectual standing, and breadth of influence, assigned before any scoring was conducted. This is an imperfect criterion: it reflects subjective judgment rather than an objective index, and it conflates several dimensions of intellectual achievement. However, it provides an external benchmark that is independent of the EC scoring process. If VRI or its component dimensions correlate with reputation, that correlation cannot be attributed to the rubric measuring its own task compliance — the rubric never saw the reputation scores, and the reputation scores were assigned without reference to the rubric.

METHOD

Sample

Thirty guests from *Conversations with Tyler* were selected to span a range of academic and intellectual fields and reputation levels. Fields represented: economics (8 guests), technology/venture (5), journalism (4), history (3), psychology (2), literature (2), forecasting (2), philosophy (1), anthropology (1), food writing (1), and finance (1). Reputation scores ranged from 5 to 10, with the following distribution: score 10 (n = 2), score 9 (n = 6), score 8 (n = 6), score 7 (n = 7), score 6 (n = 6), score 5 (n = 3). Total: 30.

All guests are English-speaking public intellectuals. This homogeneity is both a strength and a limitation: it controls for language background and interview format but restricts the generalizability of findings to this specific population.

Transcript Extraction

For each guest, 1,000--3,000 words of guest-only speech were extracted from publicly available CWT transcripts, targeting "pivot moments" — points where the host pushes the guest outside their comfort zone, forces a domain shift, or poses an unexpected challenge. This protocol captures verbal reasoning under mild cognitive load rather than rehearsed positions. All interviewer speech was removed; only guest turns were retained. Extraction was performed by multiple extractors; a formal protocol specifying word count, turn count, and pivot-moment criteria would improve replicability and is recommended for future work.

Transcript lengths varied across guests: median 1,450 words, range 549--3,012 words. Four guests fell below 1,000 words. Shorter transcripts may compress scores toward the center and are flagged as a limitation.

Scoring

The scoring procedure replicated the protocol described in the companion paper. Claude Sonnet 4 (claude-sonnet-4-20250514, Anthropic) scored each transcript blinded — identified only as "Speaker A," "Speaker B," etc. — using the full v3 EC behavioral descriptor rubric. Speaker labels were randomized independently on each of three scoring passes. Temperature was set to 0.3. The system prompt, rubric text, dimension definitions, behavioral guardrails, band descriptors, and JSON output schema were identical to the production EC assessment pipeline.

Each guest was scored in 3 independent passes. Dimension scores were averaged across passes to produce the reported scores. Total scoring passes: 90 (30 guests x 3 passes).

Reliability

Inter-pass reliability was calculated as the average spread (maximum minus minimum across 3 passes) per dimension, averaged across all 30 guests:

DIMENSION	MEAN SPREAD
Abstraction	+/-0.1
Compression	+/-0.2
Originality	+/-0.3
Conceptual Continuity	+/-0.5
Epistemic Calibration	+/-0.4

DIMENSION	MEAN SPREAD
Generative Self-Monitoring	+/-0.2
Vocabulary	+/-0.1
Syntax	+/-0.2
Overall mean	**+/-0.3**

Most dimensions achieve high stability. Conceptual Continuity shows the widest spread (+/-0.5), likely because it is most sensitive to which specific turns were extracted.

Supplementary Written-vs-Spoken Pilot

For three CWT guests — Vitalik Buterin, Marc Andreessen, and Patrick Collison — published writing samples (1,065--2,000 words) were scored blinded using the same rubric. Each written sample was scored in a single pass as "Author X" with no identifying information. Spoken scores are the 3-pass blinded averages from the main study.

External Criterion

Reputation scores were assigned before scoring, based on: (a) academic citation impact where applicable, (b) public intellectual standing as evidenced by institutional position, publication record, and media presence, and (c) breadth of influence across domains. This is a subjective, single-rater criterion. A multi-rater panel with explicit operationalized criteria would strengthen the design and is planned for future work.

RESULTS

VRI Composite vs. External Reputation

The VRI composite correlates r = 0.298 with external reputation (n = 30, p = .110). This is a weak-to-moderate positive correlation that does not reach conventional significance at alpha = .05. This attenuation is expected: all 30 guests are pre-selected by the same interviewer for intellectual distinction, compressing the VRI range to 5.8--7.5 on a 1--9 scale.

Dimension-Level Correlations

Table 1 presents Pearson correlations between each EC dimension and external reputation.

Table 1. Pearson Correlations: EC Dimensions vs. External Reputation (n = 30)

DIMENSION	R	P (TWO-TAILED)	INTERPRETATION
Generative Self-Monitoring	**0.441**	**.015**	Moderate, significant
Epistemic Calibration	**0.420**	**.021**	Moderate, significant
Compression	0.320	.085	Weak-moderate, nonsignificant
Abstraction	0.183	.333	Weak, nonsignificant (ceiling)
Vocabulary (mod.)	0.061	.750	Nil
Syntax (mod.)	0.017	.929	Nil
Originality	-0.117	.539	Weak negative, nonsignificant

DIMENSION	R	P (TWO-TAILED)	INTERPRETATION
Conceptual Continuity	-0.132	.488	Weak negative, nonsignificant

Two dimensions reach significance: Generative Self-Monitoring (r = 0.441, p = .015) and Epistemic Calibration (r = 0.420, p = .021). Together, these dimensions capture what this paper terms *generative intelligence* — the visible, real-time process of self-correction, epistemic marking, and upward revision in spontaneous speech.

Ceiling Effects

Abstraction scores cluster tightly: 26 of 30 guests scored 7 ("Principled"). This reflects sample pre-selection rather than a dimension flaw. In the MICASE study (companion paper), Abstraction discriminated well between faculty (7), graduate (6.7), and undergraduate (6) speakers.

Moderator Exclusion Validated

Vocabulary (r = 0.061) and Syntax (r = 0.017) show no correlation with reputation, confirming the design decision to exclude them from the VRI composite. In the SCOTUS domain (companion paper), these same moderators correlated r = 0.659 with a biographical attainment proxy — reflecting education and training, not reasoning quality. Their near-zero correlation in the CWT sample further supports the interpretation that they capture Gc-linked competence rather than the Gf-dominant construct EC targets.

Cross-Study Convergence: GSM Context-Sensitivity

Table 2 presents the GSM signal across three studies using the EC rubric.

Table 2. Generative Self-Monitoring Signal Across Three Speech Contexts

STUDY	CONTEXT	GSM SIGNAL
SCOTUS (companion paper)	Adversarial oral argument	Null: all advocates scored 5.0, zero spread
MICASE (companion paper)	Exploratory academic seminar	Strong differentiator: spread 2.7 (Faculty 6.0, Undergrad 3.3)
CWT (present study)	Conversational interview	Strongest predictor: r = 0.441 with reputation

This pattern is predicted by Levelt's (1983) self-monitoring model: speakers can only visibly self-correct when the speech situation permits interruption and revision. Adversarial legal argument suppresses this behavior; exploratory conversation permits it. GSM's context-sensitivity across three independent studies is evidence that the dimension measures what it claims to measure.

Supplementary: Written vs. Spoken (n = 3)

Table 3 presents dimension scores across modalities for three CWT guests.

Table 3. Written vs. Spoken Dimension Scores (n = 3 CWT Guests). Format: Written / Spoken (delta). Spoken scores are 3-pass blinded averages from the main study.

DIMENSION	BUTERIN W/S	ANDREESSEN W/S	COLLISON W/S	AVG DELTA
Abstraction	7 / 6 (+1)	7 / 7 (0)	7 / 7 (0)	+0.3
Compression	6 / 6 (0)	7 / 6 (+1)	8 / 6.3 (+1.7)	+0.9

DIMENSION	BUTERIN W/S	ANDREESSEN W/S	COLLISON W/S	AVG DELTA
Originality	6 / 7 (-1)	6 / 7 (-1)	8 / 6 (+2)	0.0
Continuity	7 / 5 (+2)	8 / 6 (+2)	8 / 7.3 (+0.7)	+1.6
Epistemic Cal	7 / 6 (+1)	4 / 6.3 (-2.3)	7 / 6.7 (+0.3)	-0.3
GSM	6 / 6 (0)	3 / 5.3 (-2.3)	7 / 5.3 (+1.7)	-0.2
Vocabulary	6 / 7 (-1)	7 / 7 (0)	7 / 7 (0)	-0.3
Syntax	7 / 6 (+1)	7 / 6 (+1)	7 / 6 (+1)	+1.0

Three patterns emerge. First, **Syntax consistently increases in writing** (average delta +1.0; all three guests show +1), reflecting the editorial revision that polished prose undergoes. Second, **Conceptual Continuity consistently increases in writing** (average delta +1.6), reflecting the cumulative structure that editing creates. Third, **Epistemic Calibration and GSM are person-by-medium interactions**: Andreessen's written Epistemic Calibration drops from 6.3 (spoken) to 4 (written) and his GSM drops from 5.3 to 3 — his manifesto is declarative, with no hedging and no self-correction. By contrast, Collison's GSM increases from 5.3 (spoken) to 7 (written), and Buterin's Epistemic Calibration increases from 6 to 7. These are writers whose prose preserves — and in Collison's case amplifies — their epistemic habits. The spontaneous speech condition surfaces epistemic complexity that some writers edit out of their prose, and preserves it for others.

DISCUSSION

The Central Finding

Within a pre-selected sample of 30 high-ability conversational speakers, Generative Self-Monitoring (r = 0.441) and Epistemic Calibration (r = 0.420) are the only EC dimensions that significantly predict external intellectual reputation. Together they capture a mode of spoken intelligence this paper terms *generative* — characterized by visible real-time self-correction, spontaneous epistemic marking, and upward revision. This contrasts with *performative* intelligence — characterized by confident, compressed, abstract delivery — which ceilings out in this sample and does not discriminate.

The distinction is not between smart and not-smart. All 30 guests are intellectually accomplished. The distinction is between speakers who *display the process of thinking* and speakers who *display the products of thought*. The EC rubric, with its dimension-level granularity, captures this distinction where a unidimensional composite cannot.

Why VRI Underperforms Its Dimensions

The VRI composite (r = 0.298) underperforms both GSM (r = 0.441) and Epistemic Calibration (r = 0.420). This is because the composite includes four dimensions that do not predict reputation in this sample: Abstraction (ceiling effect), Compression (moderate), Originality (slightly negative), and Conceptual Continuity (slightly negative). These dimensions measure real cognitive operations — they discriminate well in other contexts (companion paper) — but they do not track reputation in a sample where everyone is already abstract, compressed, and articulate.

This finding has a practical implication: within high-ability populations, the dimension profile is more informative than the VRI composite. A user whose GSM and Epistemic Calibration are both 7+ is

displaying reasoning behaviors associated with the highest-reputation thinkers in this sample, regardless of their VRI.

Disciplinary Reasoning Signatures

Without being told guests' fields, the rubric produces dimension profiles that cluster by discipline. Philosophers show high GSM, high Epistemic Calibration, and high Continuity — training in real-time argumentation is visible in the speech. Economists show high Abstraction with moderate scores elsewhere — framework-driven reasoning. Technology founders show high Abstraction and Compression but lower GSM and Continuity — performative, conclusory delivery. Writers and journalists show higher Continuity and Originality — narrative coherence rather than epistemic self-correction. These clusters were not hypothesized in advance, were identified post hoc by inspecting the scored data, and should be treated as unplanned exploratory observations rather than confirmatory findings. However, their theoretical coherence — and their alignment with what is known about how disciplinary training shapes discourse practices — suggests the rubric may be detecting something real about the relationship between professional formation and spontaneous verbal reasoning. Confirmatory testing with a priori domain-based hypotheses and adequate within-domain sample sizes is required before these patterns can be treated as established.

The "I Don't Know" Hypothesis

Guests who score highest on Epistemic Calibration are those who voluntarily mark the boundary of their knowledge — saying "I don't know," "I'm not sure about this," or "that's where my confidence drops" without being prompted. Guests who score lower treat all claims with uniform confidence. This pattern suggests that the most measurable form of intellectual sophistication in conversational speech may be the spontaneous differentiation of one's own certainty — a metacognitive operation that is difficult to perform without the underlying capacity.

LIMITATIONS

- **Reputation scores are subjective.** A single rater assigned all scores before scoring. A multi-rater panel with explicit, operationalized criteria would strengthen the external criterion.
- **Transcript extraction was not standardized.** Different extractors selected different turns. This introduces uncontrolled variance: extractors who selected more cognitively demanding moments may have inflated scores relative to extractors who selected more routine passages. The direction of this bias is unpredictable across guests. A formal protocol specifying word count, turn count, and pivot-moment criteria is needed for replicability.
- **Sample is homogeneous.** All guests are English-speaking public intellectuals from the same podcast. Generalizability to other populations, languages, or conversational contexts requires further study.
- **Short transcripts for some guests.** Four guests had fewer than 1,000 words of extracted speech. Shorter samples may compress scores toward the center and attenuate correlations.
- **Single scorer.** Only Claude Sonnet 4 was used. Cross-model replication, as implemented in the SCOTUS studies (companion paper), would strengthen confidence.
- **No causal claims.** The correlations between EC dimensions and reputation are observational. Reputation may cause certain reasoning behaviors; reasoning behaviors may cause reputation; or both may be caused by a third factor.

- **The written-vs-spoken pilot is underpowered.** Three guests, single-pass written scoring. Patterns are theoretically interpretable but require replication with larger n and matched pass counts.
- **The elicitation context is not the EC assessment.** This study validates the scoring rubric applied to naturalistic conversational speech, not the standardized EC prompt set. Prompt validation is a separate study currently planned.

CONCLUSION

In a sample of 30 high-ability conversational speakers, the EC rubric detects meaningful dimensional variation that correlates with external intellectual reputation. The finding is dimension-specific: Generative Self-Monitoring and Epistemic Calibration — the two dimensions that capture visible, real-time cognitive process rather than accumulated knowledge — are the only significant predictors. This result converges with prior construct validity evidence (companion paper): GSM is null in adversarial speech, strong in exploratory speech, and strongest in conversational speech, a pattern predicted by Levelt's (1983) self-monitoring model and consistent with the interpretation that EC measures reasoning as performed.

The written-vs-spoken pilot adds a convergent validity signal: dimensions that measure the person (Vocabulary, Abstraction) are stable across modalities, while dimensions that measure the process (GSM, Continuity) shift in predictable, person-specific ways. Spontaneous speech is not a limitation of the EC design — it is the medium in which generative intelligence is most visible and least manageable.

What distinguishes high-reputation thinkers in spontaneous conversational speech is not what they know — it is how visibly they think.

REFERENCES

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy.* Academic Press.

Construct Validity of the Verbal Reasoning Index: Evidence from Naturalistic Speech Corpora. Companion paper, available at expressivecognition.org.

Flavell, J. H. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*(10), 906--911. https://doi.org/10.1037/0003-066X.34.10.906

Guilford, J. P. (1967). *The nature of human intelligence.* McGraw-Hill.

Kintsch, W. (1974). *The representation of meaning in memory.* Erlbaum.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*(1), 41--104. https://doi.org/10.1016/0010-0277(83)90026-4

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* MIT Press.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741--749. https://doi.org/10.1037/0003-066X.50.9.741

Simonton, D. K. (1999). *Origins of genius: Darwinian perspectives on creativity.* Oxford University Press.

Simpson, R., Briggs, S., Ovens, J., & Swales, J. M. (2002). *The Michigan corpus of academic spoken English.* The Regents of the University of Michigan.

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., & Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. *JAMA, 275*(7), 528--532. https://doi.org/10.1001/jama.1996.03530310034031

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* Academic Press.

Vygotsky, L. S. (1962). *Thought and language.* MIT Press.

APPENDIX A: FULL SAMPLE DATA

Table A1. Complete dimension scores for all 30 CWT guests. Scores are 3-pass blinded averages (Claude Sonnet 4, v3 rubric). Reputation scores (1--10) were assigned before scoring based on citation impact, public intellectual standing, and breadth of influence. These scores reflect an assessment of intellectual reputation in public discourse; they are not measures of intelligence, character, or professional competence.

GUEST	FIELD	REP	ABSTR	COMPR	ORIG	CONT	EPIST	GSM	VOCAB	SYN
Daron Acemoglu	Economics	10	7	6	6	7	7.7	6.3	7	6
Larry Summers	Economics	10	7	7.3	6	6.3	7.3	6	8	7
Philip Tetlock	Forecasting	9	7.3	6.3	7	6.7	8	6	7.3	6.3
Cass Sunstein	Law/Policy	9	7	6	7	6	8	6.7	7	6
Raj Chetty	Economics	9	7	6	6.7	7.7	7	6	7	6.7
Steven Pinker	Cognitive Science	9	7	6	6	7	7.3	6	7	6
Paul Krugman	Economics	9	7	6	6	5	8	7	7	6
Sam Altman	Tech	9	7	6	7	6	7	6	6.3	5.3
Alison Gopnik	Psychology	8	8	7	7	7.3	8	6.7	7.3	6
Joseph Henrich	Anthropology	8	7	6	7	7.7	7	6	7	6
Jonathan Haidt	Psychology	8	7	6	7	8	7	6	7	6
Malcolm Gladwell	Journalism	8	7	6	7.3	6.7	6.7	6	7	6.7
Margaret Atwood	Literature	8	7	6	7.3	5.7	7.3	6	7	6

GUEST	FIELD	REP	ABSTR	COMPR	ORIG	CONT	EPIST	GSM	VOCAB	SYN
Nassim Nicholas Taleb	Finance/Philosophy	8	6.3	7.3	7.3	5	4.3	7.7	7	6
Agnes Callard	Philosophy	7	8	7	7	8	8	7	8	7
David Brooks	Journalism	7	7	6	7	8	7	6	7	6
Larissa MacFarquhar	Journalism	7	7	6	7	7.7	6.7	6	7	6.7
Nate Silver	Forecasting	7	7	6	6	7	7.3	6	7	6
Patrick Collison	Tech	7	7	6.3	6	7.3	6.7	5.3	7	6
Marc Andreessen	Tech/VC	7	7	6	7	6	6.3	5.3	7	6
Ezra Klein	Journalism	7	7	6	6	5	6.7	5.7	7	6
Fuchsia Dunlop	Food Writing	6	7	6	7	8	7	6	8	7
Russ Roberts	Economics	6	7	6	7	8	6.7	6.3	7	6.3
Sam Bankman-Fried	Finance/Tech	6	7	6	7	6	7	6	7	6
Paul Gillingham	History	6	7	6	6	7	6	5	7	6
Vitalik Buterin	Tech/Crypto	6	6	6	7	5	6	6	7	6
Emily St. John Mandel	Literature	6	6	5	7	6	6	5	6	6
Mark Koyama	History	5	7	6	6.3	7.3	7.3	6	7	6
Noel Johnson	History	5	7	6	6.3	7.3	6.7	5.3	7.3	6.3
Daniel Gross	Tech/Investing	5	7	6	7	6.3	6	5.3	7	6

Note. Abstr = Abstraction; Compr = Compression; Orig = Originality; Cont = Conceptual Continuity; Epist = Epistemic Calibration; GSM = Generative Self-Monitoring; Vocab = Vocabulary (moderator); Syntax = Syntactic Control (moderator). VRI = Verbal Reasoning Index composite (core dimensions only, weighted). All scores on 1--9 scale. Sorted by reputation score (descending), then VRI (descending) within reputation tier.