Expressive Cognition: Measuring How People Think in Real Time

A White Paper on the Verbal Reasoning Index

The Problem No Existing Test Solves

There is a moment that everyone who works with people recognizes. It happens in interviews, in meetings, in clinical encounters, in classrooms. Someone is asked a question they haven't prepared for, and you watch what happens next.

Some people arrive somewhere unexpected. They hold the question from multiple angles simultaneously, notice something the question didn't anticipate, and say something that makes you think differently about the problem. Others produce the answer the question most obviously calls for — accurate, fluent, unremarkable. Still others fragment. They circle back, lose the thread, arrive somewhere close to where they started.

These differences are real. They predict things that matter: how someone will perform in a novel role, how quickly they learn in unfamiliar domains, how they reason under pressure when no prepared answer exists. And yet no standard assessment captures them directly from speech.

Existing verbal ability tests — the GRE Verbal, IELTS, TOEFL, WAIS-IV Verbal Comprehension — measure accumulated linguistic knowledge: vocabulary breadth, reading comprehension, the ability to manipulate learned information. These are valuable measures. But they primarily reflect what someone has absorbed over a lifetime of education and reading. They are less sensitive to what someone does with that knowledge in real time, under novel conditions, when the question arrives before the answer has been prepared.

Intelligence tests measure reasoning capacity — but through abstract, non-verbal stimuli in controlled environments, administered by trained examiners. They are not designed to capture reasoning as it actually occurs in human communication: through language, spontaneously, in response to the unexpected.

Communication assessments — speaking rubrics, interview scoring guides, presentation evaluations — measure delivery, fluency, and organizational skill. They score how someone communicates, not the cognitive quality of what is being communicated.

Expressive Cognition occupies the space none of these instruments address: the direct measurement of reasoning quality in spontaneous speech, from a short spoken response to a novel question, scored without a human examiner.

What EC Measures

EC's Verbal Reasoning Index (VRI) measures the cognitive quality of spontaneous verbal production under novel conditions.

The key terms in that definition each carry specific meaning.

Cognitive quality — not fluency, not grammatical accuracy, not vocabulary level. Those surface features of speech are captured as context but are not the primary signal. The primary signal is what is happening cognitively: whether ideas cohere, whether thought develops, whether something new arrives.

Spontaneous — not prepared, not rehearsed. The speaker may know they are being recorded. They do not know what the question will be. What surfaces in the response is the unguarded cognitive signature — the shape of thinking before it has been shaped into performance.

Novel conditions — the elicitation is designed to prevent pattern-matching to a familiar genre. A question the speaker cannot have prepared for requires genuine real-time reasoning. The assessment is sensitive to this because it is scoring the process of meaning-making, not the product of preparation.

From speech — not from writing, not from a structured test, not from an interview where a human examiner is present to probe and scaffold. A short spoken response, typically 60–90 seconds, is sufficient to detect the cognitive signature. The measure is scalable in a way

that clinical assessment is not.

The Six Dimensions

EC scores six core dimensions, each targeting a specific cognitive operation that is observable in spontaneous speech. The dimensions are scored 1–9 using behavioral descriptor bands grounded in established psychometric and linguistic frameworks.

1. Abstraction

What level does the speaker naturally think at?

When a person encounters a novel question, they can stay close to the specific case — describing instances, naming examples, staying at the surface of what the question asks. Or they can reach for the category, the principle, the framework that the specific case belongs to.

This is not about using abstract-sounding words. Saying "fundamentally" or "at its core" is not abstraction. Deriving a principle from an example is. Abstraction measures where the speaker spontaneously lands on the concrete-to-principled continuum — not where they can reach when prompted, but where they naturally go.

Theoretical grounding: SOLO Taxonomy (Biggs & Collis, 1982); Vygotsky's concept formation levels

2. Compression

How much can the speaker hold simultaneously?

Propositional density — the ratio of ideas to words — is one of the most robust cognitive markers in spontaneous speech. Low compression means circling, restating, buying time with words. High compression means ideas arrive already integrated: a single clause doing the work of three sentences, no redundancy, the speaker consistently ahead of the sentence they are constructing.

This is not brevity. A highly compressed response can be long. The measure is how much cognitive content arrives per unit of language.

Theoretical grounding: Kintsch's propositional analysis (1974); Snowdon et al. (1996) idea density as cognitive predictor

3. Originality

Does anything genuinely new arrive?

The expected response to a novel question is the most culturally available answer — the thing anyone would say. Originality scores whether the speaker arrives somewhere else: a non-obvious frame, an analogy from an unexpected domain, a reversal of the question's implicit assumption.

The criterion is aptness plus novelty. A surprising framing that illuminates the problem is high Originality. A merely unconventional response that does not advance understanding is not.

Originality is the dimension most resistant to preparation and most dependent on what happens cognitively in the moment of responding. Across every validity study conducted so far, it is the most robust differentiator of reasoning quality at all ability levels.

Theoretical grounding: Guilford's divergent production; Finke et al.'s geneplore model (1992)

4. Conceptual Continuity

Does thought build, or does it fragment?

The question is not whether the response is logically organized — logical organization is trainable. The question is whether ideas accumulate: whether each move derives from and advances what came before, whether the response arrives somewhere the opening did not anticipate, whether thinking visibly develops across the response.

Fragmentation — topic shifts without transition, ideas that do not follow from previous ones, responses that end where they started — is a reliable signal of working memory load and the limits of real-time cognitive integration.

Theoretical grounding: Halliday & Hasan's cohesion theory (1976); van Dijk & Kintsch's situation model (1983)

5. Epistemic Calibration

Does the speaker know what they know?

Calibration measures whether the speaker spontaneously distinguishes what they are certain of from what they are inferring — without being prompted, without being challenged. High calibration is not excessive hedging. It is differentiated confidence: the speaker marks the difference between a strong claim and a tentative one, between direct knowledge and inference.

This distinction — spontaneous rather than trained — matters. Professional hedging (learned in law school, corporate communication, or clinical training) looks similar on the surface but reflects a different cognitive operation than genuine epistemic monitoring. EC's scoring criteria distinguish these.

Theoretical grounding: King & Kitchener's Reflective Judgment Model (1994)

6. Generative Self-Monitoring

Does the speaker improve in real time?

When a speaker says something, notices it is not quite right, and says something better — not just correcting a word, but reformulating an idea at a higher level — that is the positive signal of active reasoning becoming aware of itself. It is evidence that thinking is happening, not just retrieval.

This dimension measures whether revision moves ideas upward. It does not penalize disfluency — filled pauses and false starts are processing signals, not reasoning failures. The signal is targeted conceptual revision: catching an imprecision and improving it, not just cleaning up language.

Theoretical grounding: Levelt's self-monitoring model (1989); Flavell's metacognition theory (1979)

What EC Does Not Measure

Being explicit about what EC does not measure is as important as what it does.

EC does not measure linguistic competence as the primary signal. Vocabulary breadth and syntactic control are scored as contextual moderators — reported alongside the VRI but not included in the composite. This is a deliberate decision. These features reflect educational exposure and language background more than cognitive quality in the moment. They are not the construct.

EC does not measure preparation or domain knowledge. A speaker who has extensively prepared a topic can produce a response that sounds sophisticated without any real-time reasoning occurring. EC's elicitation design — novel prompts, unknown in advance — is specifically intended to prevent this. But EC is not impervious to domain familiarity; a speaker discussing a topic they know deeply will naturally score higher on some dimensions. This is acknowledged in the scoring methodology and in the contextual moderation framework.

EC does not diagnose anything. It is not a clinical instrument, not a substitute for neuropsychological evaluation, and not a measure of intelligence in the general sense. It measures a specific cognitive performance under specific conditions.

EC is not a verbal IQ test. The relationship between EC scores and standard verbal IQ measures is real — they share variance because accumulated linguistic knowledge supports real-time reasoning — but it is not identity. EC captures something that verbal IQ tests, by design, do not: the quality of reasoning as it occurs in real-time speech, before preparation is possible. Empirically, the expected correlation between EC scores and verbal IQ measures is meaningful but not strong enough to make the measures redundant. They are measuring related but distinct things.

The Evidence

EC's validity program applies the same methodological framework that governs any serious psychometric instrument: known-groups designs, blinded scoring, multi-pass reliability testing, and cross-model replication.

Known-Groups Study 1: Supreme Court Oral Arguments

EC's scoring rubric was applied to oral argument transcripts from three attorneys representing three tiers of SCOTUS experience: Paul Clement (elite, ~92 arguments, 75.8% win rate), Kannon Shanmugam (experienced, 15+ arguments), and Richard Dearing (first-time advocate). Experience tiers were externally established — not inferred from EC scores.

The study employed a 2×2 design crossing two scoring models (Claude Sonnet 4, GPT-4o) and two framing conditions (unblinded, blinded), with three passes per condition to measure scoring reliability.

Results: Claude Sonnet 4, scoring blindly across three independent passes with randomized speaker labels, consistently ranked the attorneys in the predicted order (Clement > Shanmugam > Dearing). The separation was small but perfectly stable — identical scores on every pass. GPT-4o maintained the correct rank order when given speaker identity but produced flat scores when blinded, revealing a large expectation effect.

Dimension-level findings were theoretically interpretable. Originality was the only dimension where both models, scoring blindly, agreed on the full three-way rank order. Epistemic Calibration — which scores differentiated confidence rather than trained hedging — showed no differentiation among Supreme Court advocates, consistent with the expectation that all three attorneys are trained to manage epistemic claims before the court and this skill saturates at the elite level.

Generative Self-Monitoring was null across all attorneys and all conditions: all three scored identically. This was interpretable as a property of the elicitation context — adversarial oral argument does not afford real-time revision because justices interrupt before speakers can refine — rather than a failure of the dimension. Study 2 confirmed this interpretation.

This study demonstrates that EC's construct generalizes beyond its standardized prompt conditions to naturally occurring high-stakes verbal reasoning.

Known-Groups Study 2: Academic Discourse (MICASE)

EC's rubric was applied to spontaneous speech from the Michigan Corpus of Academic Spoken English, selecting speakers at three academic role tiers: undergraduate student, graduate student, and faculty. Same scoring design.

The study used a corrected design in which all three speakers were performing the same cognitive task: presenting and defending their own ideas under questioning. The faculty member and graduate student were in the same philosophy seminar — the same room, same topic, same session — with the faculty member defending his own theory and the graduate student challenging it with his own position. The undergraduate was in a separate seminar defending her own budget analysis.

Results: The full predicted rank order held in the blinded condition with clean 1-point gaps between tiers (Faculty 6.7, Graduate 5.7, Undergraduate 4.5). Generative Self-Monitoring — which showed no variation in the SCOTUS study, where adversarial argument does not afford revision time — emerged as the strongest differentiating dimension here, with a 2.7-point spread from faculty to undergraduate. The seminar format, unlike adversarial oral argument, allows speakers to catch and reformulate in real time. The undergraduate's surface-level repairs were clearly distinguishable from the faculty's conceptual revisions.

The faculty member was the only speaker in either study to reach the "Generative" band on Originality — inventing philosophical arguments in real time rather than reframing existing ones.

Average dimension-level spread across three blinded passes: ±0.1. Near-deterministic scoring.

Preliminary Ecological Validity: Conversations with Tyler (Pilot)

A pilot study (n=11) applied EC's rubric to guest speech from the podcast *Conversations with Tyler*, hosted by economist Tyler Cowen. Guests spanned multiple domains (economics, cognitive science, law, journalism, technology) and externally assigned reputation tiers (6–

10 on a composite scale combining citations, awards, institutional standing, and publication impact).

A critical methodological finding emerged from this study: **scores are highly sensitive to which moments of speech are scored.** When guest turns were extracted without regard to elicitation quality — including sections where guests explained their published work or recited familiar material — the correlation between VRI and reputation was weak (r = 0.298). When extraction targeted only "pivot moments" — sections where the host pushed guests off their prepared material with unexpected questions, domain shifts, or challenges to their thesis — the correlation rose to moderate (r = 0.395), driven primarily by Epistemic Calibration (r = 0.430) and Generative Self-Monitoring (r = 0.375).

Abstraction and Compression showed near-zero correlation with reputation in this sample — not because they are poor dimensions, but because all of Cowen's guests reason at the Principled and Dense levels. These dimensions ceiling out in a high-ability sample. The signal that differentiates within a compressed ability range is whether speakers spontaneously monitor the certainty of their own claims and whether they revise their formulations upward in real time.

This study is preliminary — 11 guests is insufficient for stable correlational estimates, and the reputation scores involve subjective judgment. It is reported here because the methodological finding about elicitation quality has direct implications for EC's scoring protocol: **the construct lives in the moments of genuine real-time reasoning, not in the moments of prepared exposition.**

Scoring Reliability

Across all studies, Claude Sonnet 4 produced scoring that is approximately twice as reliable as GPT-4o in blinded conditions. On most dimensions, it achieves identical scores across multiple passes of the same text — a level of reliability that exceeds many human inter-rater reliability benchmarks in published psychometric research.

A Note on What Has and Has Not Been Validated

The validity evidence described above was collected on **naturally occurring speech corpora** — Supreme Court transcripts, academic seminar recordings, and podcast interviews. These are not the same as EC's standardized elicitation, which uses its own set of novel spoken prompts designed to elicit different types of reasoning (explanation, analogy, compression, argument, abstract reasoning).

EC's standardized prompt set has not yet been independently validated. The prompts were designed based on the psychometric principles described in this paper, and they are the elicitation that users encounter on expressivecognition.org. But the studies reported here do not test those specific prompts — they test the scoring rubric applied to speech elicited under different conditions.

This is an important distinction. The evidence establishes that EC's scoring rubric detects real verbal reasoning differences in spontaneous speech across multiple contexts. It does not yet establish that EC's own prompts elicit the construct as effectively as the naturalistic conditions in the validation studies. Establishing this requires a dedicated study comparing EC prompt-elicited speech to alternative elicitations in the same speakers — a study that is planned but not yet conducted.

Until that study is complete, the claim is: **EC's rubric measures what it claims to measure when applied to speech that contains genuine real-time reasoning. Whether EC's own prompts reliably produce such speech is a separate empirical question that has not yet been answered.**

This is stated plainly because the distinction matters. A scoring rubric can be valid while a particular elicitation is suboptimal. The planned normative study on Prolific will be the first test of the standardized prompts under controlled conditions.

Who EC Is For

High-stakes hiring and talent identification. The outcomes that predict success in complex professional roles — adaptive reasoning under pressure, learning rate in novel domains, performance when no prepared answer exists — are better captured by EC than by credentials, GPA, or verbal ability tests alone. EC provides a signal on the cognitive quality of real-time reasoning, in a format that feels like a conversation rather than a test.

Language learning assessment. The plateau between advanced and proficient language use is often a reasoning plateau, not a linguistic one. A learner who has reached the limits of what grammar instruction and vocabulary study can provide may need a different kind of intervention — one targeting the capacity for real-time meaning-making. EC can identify this.

Cognitive monitoring. Spontaneous verbal cognition is one of the earliest behavioral systems affected by cognitive decline. EC's sensitivity to the dimensions that degrade in mild cognitive impairment — Compression, Conceptual Continuity, Generative Self-Monitoring — makes it a candidate for longitudinal monitoring applications. This remains a research direction rather than a current clinical application.

Research. EC provides a methodology — a theoretically grounded, reliably scored rubric for measuring cognitive quality in spontaneous speech — that can be applied to any corpus of transcribed speech. No comparable instrument exists.

What Comes Next

EC's validity program is ongoing. The studies described above are the first in a planned sequence:

Standardized elicitation validation. A study comparing VRI scores from EC's own prompts to scores from naturalistic elicitation in the same speakers. This is the study that tests whether the product elicits the construct as well as the validation corpora do.

A clinical validity study using the DementiaBank Pitt Corpus will examine whether EC scores track the HC → MCI → AD gradient in Cookie Theft picture description, with the MMSE as an external criterion. This tests whether EC's dimensions are sensitive to cognitive decline at clinically established levels.

A normative and convergent validity study using Prolific will establish VRI norms across a healthy adult sample and test whether EC scores correlate with established fluid reasoning measures (Raven's Progressive Matrices, WAIS-IV Fluid Reasoning Index) above and beyond verbal IQ. This is the study that tests EC's central claim: that it measures something real that existing instruments miss.

A Note on Transparency

EC is a novel instrument in active development. The validity program described here represents current evidence, not a completed validation portfolio. The claims made on this basis are calibrated accordingly.

What can be said with confidence: EC's scoring rubric detects real verbal reasoning differences in naturally occurring speech, across multiple ability levels, with scoring reliability that meets or exceeds published psychometric standards.

What remains to be established: whether EC's own standardized prompts elicit the construct as effectively as naturalistic conditions; normative data across a general population; convergent validity with established cognitive measures; and clinical sensitivity across impairment levels.

These are the studies being conducted. The methodology is public. The rubric is open. Researchers who wish to apply EC's framework to their own corpora or build on this work are encouraged to do so.

Technical Summary

Feature	Specification
Scoring dimensions	6 core (VRI) + 2 moderators
Scale	1–9 per dimension, behavioral descriptor bands
Scoring model	Claude Sonnet 4
Elicitation (product)	5 novel spoken prompts, ~10 minutes total

Elicitation (validation)	SCOTUS transcripts, MICASE academic speech, CWT podcast speech
Reliability (avg spread, blinded)	±0.1–0.2 across passes
Known-groups validation	3 studies across legal, academic, and intellectual discourse
Standardized prompt validation	Planned (not yet conducted)
In progress	DementiaBank clinical study, Prolific normative study

---

Expressive Cognition is available at expressivecognition.org. The scoring rubric and validity study documentation are available for research use.