

Construct Validity of the Verbal Reasoning Index: Evidence from Naturalistic Speech Corpora

ABSTRACT

The Expressive Cognition (EC) Verbal Reasoning Index scores spontaneous speech across six core dimensions and two moderators using an AI-scored behavioral descriptor matrix. This paper tests a single claim: the EC rubric detects known differences in verbal reasoning quality in the predicted direction, across two independent naturalistic speech corpora, at different points on the ability distribution. Study 1 applies the rubric to Supreme Court oral arguments from three attorneys at distinct experience tiers. Study 2 applies it to academic seminar speech from faculty, graduate, and undergraduate speakers. Both studies use blinded scoring with multiple passes and cross-model replication. The rubric correctly ranks speakers in both corpora, with high inter-pass reliability (mean spread +/-0.2 on a 1--9 scale). Two emergent findings are theoretically significant: Generative Self-Monitoring is null in adversarial legal argument but strongly differentiates in exploratory academic discussion, and Originality inverts the expected hierarchy when the cognitive task performed by a lower-tier speaker is more generative than the task performed by a higher-tier speaker. These results establish that EC scores verbal reasoning as performed, not speaker credential, and that the construct generalizes beyond EC's own elicitation protocol.

INTRODUCTION

The EC scoring rubric was designed to evaluate verbal reasoning in 60--90 second spoken responses to standardized prompts. If the rubric measures a real construct — not merely compliance with its own task format — it should also detect reasoning quality differences in naturally occurring speech that was produced without any EC involvement. This is a stronger test than internal consistency or test-retest reliability because it requires the construct to travel across contexts.

We select two corpora that provide known-groups contrasts:

Study 1: Supreme Court Oral Arguments. Three attorneys at distinct tiers of experience and reputation argue before the Court. The speech is adversarial, high-stakes, and produced under extreme time pressure. If EC detects the predicted rank order (elite > experienced > first-time), the construct generalizes to professional legal discourse.

Study 2: MICASE Academic Seminars. Three speakers at distinct academic levels — faculty, graduate student, undergraduate — participate in university seminars. The speech is exploratory, collaborative, and produced under moderate cognitive load. If EC detects the predicted rank order (faculty > graduate > undergraduate), the construct generalizes to the full academic ability range.

Together, these studies test generalizability at both ends of the distribution and in two fundamentally different speech contexts — one adversarial, one collaborative. The claim is deliberately narrow: we are testing whether the rubric *detects* known differences, not whether it predicts external outcomes (that is the work of a separate ecological validity study).

METHOD

The EC Scoring Rubric

The EC rubric scores speech on six core dimensions aggregated into the Verbal Reasoning Index (VRI) and two moderator dimensions reported separately:

Core dimensions (included in VRI):

DIMENSION	CONSTRUCT	BANDS (1--9)
Abstraction	Where on the concrete-to-principled continuum	Concrete / Classifying / Relational / Principled
Compression	Propositional density per unit of language	Expansive / Sequential / Dense / Packed
Originality	Genuinely unexpected framing that illuminates	Conventional / Inflected / Reframed / Generative
Conceptual Continuity	Whether ideas build cumulatively	Fragmented / Listed / Connected / Cumulative
Epistemic Calibration	Spontaneous differentiation of confidence across claims	Undifferentiated / Occasionally Hedged / Differentiated / Reflexively Calibrated
Generative Self-Monitoring	Real-time self-correction and upward revision	Unreflective / Surface Repair / Conceptual Revision / Generative Refinement

Moderator dimensions (not in VRI):

DIMENSION	CONSTRUCT	BANDS (1--9)
Vocabulary	Lexical precision and diversity	Basic / Functional / Precise / Surgical
Syntax	Sentence-level structural control	Simple / Basic / Controlled / Fluent

Each dimension includes explicit "what it is / what it is NOT" guardrails. The scorer matches sustained performance — not peak moments — to the highest band consistently demonstrated. One evidence quote is required per dimension.

VRI weights: Abstraction (0.18), Epistemic Calibration (0.18), Compression (0.16), Originality (0.16), Conceptual Continuity (0.16), Generative Self-Monitoring (0.16). The weighted average is mapped to a scale centered at 100 with SD=15, reported as a range (e.g., 105--113) to reflect single-session variability.

LLM-as-Scorer Procedure

All scoring was performed by large language models acting as automated raters. The procedure was:

- **Prompt construction.** Each transcript was embedded in a scoring prompt containing the full descriptor matrix, band definitions, guardrails, and a JSON output schema. The prompt included no information about the speaker's identity, credentials, or institutional affiliation in the blinded condition. In the unblinded condition, the speaker's name, role, and context were provided.
- **Model selection.** Two models were used: Claude Sonnet 4 (claude-sonnet-4-20250514, Anthropic) and GPT-4o (OpenAI). Claude served as the primary scorer based on prior rubric development testing that showed it produced the most stable and calibrated scores. GPT-4o served as an independent replication scorer.
- **Blinding protocol.** In the blinded condition, speakers were assigned randomized labels (Speaker A, Speaker B, Speaker C) that were reshuffled on each pass. The scorer received no information beyond "Spontaneous spoken responses. No other information." In the unblinded condition, the speaker's full name, professional description, and case context were provided.
- **Multi-pass scoring.** Each speaker was scored in 3 independent passes per condition. Passes used independently randomized speaker labels. Temperature was set to 0.3 to permit some scoring variation while constraining hallucination.
- **Reliability calculation.** Inter-pass reliability was calculated as the average spread (maximum score minus minimum score across 3 passes) for each dimension, averaged across speakers. A spread of 0 indicates identical scores on all passes; a spread of 1 indicates the model varied by one point on the 1--9 scale.
- **JSON parsing.** Model outputs were parsed as JSON with fallback cleaning for markdown fences and preamble text. Malformed responses were retried up to twice.

Study 1: SCOTUS Oral Arguments

Speakers. Three Supreme Court advocates at distinct experience tiers:

SPEAKER	TIER	ARGUMENTS	NOTABLE CASE	WORDS
Paul Clement	Elite	~92	<i>NetChoice v. Paxton</i> (2024)	1,141
Kannon Shanmugam	Experienced	15+	<i>Seila Law v. CFPB</i> (2020)	950

SPEAKER	TIER	ARGUMENTS	NOTABLE CASE	WORDS
Richard Dearing	First-time	1	<i>NY Rifle & Pistol v. NYC</i> (2019)	3,886

Predicted rank order: Clement > Shanmugam > Dearing on VRI composite.

Transcript source. Official Supreme Court oral argument transcripts. Advocate-only speech was extracted (all questions from justices removed). Transcripts were drawn from cases argued between 2019 and 2024.

Scoring. Both Claude Sonnet 4 and GPT-4o scored all three speakers in both unblinded and blinded conditions, 3 passes each. Total: 36 scoring passes (3 speakers x 2 conditions x 2 models x 3 passes).

Study 2: MICASE Academic Seminars

Speakers. Three speakers from the Michigan Corpus of Academic Spoken English (MICASE), spanning the academic continuum:

SPEAKER	LEVEL	SEMINAR	ROLE IN TRANSCRIPT	WORDS
Faculty (S1)	Professor	SEM475 Philosophy	Presenting own paper on phenomenal illusions	1,973
Graduate (S2)	PhD student	SEM475 Philosophy	Challenging faculty member's theory	2,430
Undergraduate (S4)	Undergrad	SEM495 Politics of Higher Ed	Presenting own budget analysis	746

Predicted rank order: Faculty > Graduate > Undergraduate on VRI composite.

Task confound (documented). Faculty and Graduate participated in the same seminar session (SEM475), but the Faculty member was presenting original theoretical work while the Graduate student was critiquing it. The Undergraduate was in a different seminar (SEM495), presenting an empirical analysis of university budgets. These are different cognitive tasks — presenting theory, critiquing theory, and presenting data — and the rubric may respond to task type as well as speaker ability. We treat this as informative rather than confounding (see Discussion).

Scoring. Claude Sonnet 4 only (based on its superior blinded performance in Study 1). Unblinded and blinded conditions, 3 passes each. Total: 18 scoring passes (3 speakers x 2 conditions x 3 passes).

RESULTS

Study 1: SCOTUS — VRI Composite

CONDITION	CLEMENT	SHANMUGAM	DEARING	PREDICTED ORDER?
Claude Unblinded	6.5 (107--115)	6.0 (104--112)	5.8 (102--110)	Yes
Claude Blinded	6.2 (105--113)	5.9 (103--111)	5.9 (102--110)	Yes (Clement top)
GPT-4o Unblinded	6.0 (104--112)	5.8 (102--110)	5.2 (98--106)	Yes
GPT-4o Blinded	5.5 (100--108)	5.6 (100--108)	5.2 (98--106)	Partial (0.1 swap at top)

Claude maintains the full predicted rank order in all conditions, including blinded. GPT-4o preserves the rank order unblinded but swaps Clement and Shanmugam by 0.1 when blinded — within the noise floor. Both models separate Dearing from the top two in every condition.

Expectation effects. Blinding reduces Clement's VRI by 0.3 (Claude) and 0.5 (GPT-4o). Shanmugam's changes by 0.1 or less. Dearing's is essentially unchanged. The v3 descriptor-matrix rubric produces smaller expectation effects than earlier rubric versions (which showed 0.6+ shifts in some conditions).

Study 1: SCOTUS — Dimension-Level Scores (Claude Blinded, 3-Pass Average)

DIMENSION	CLEMENT	SHANMUGAM	DEARING	SPREAD
Abstraction	7.0	7.0	6.7	0.3
Compression	6.0	6.0	6.3	0.3
Originality	6.0	4.3	4.0	2.0
Conceptual Continuity	7.0	7.0	6.7	0.3
Epistemic Calibration	6.0	6.0	6.3	0.3
Gen. Self-Monitoring	5.0	5.0	5.0	0.0
Vocabulary	7.0	7.0	7.0	0.0
Syntax	6.0	6.0	6.0	0.0

Originality is the sharpest differentiator. It is the only core dimension where Clement receives a different band label from the other two even when blinded: "Reframed" vs. "Inflected." The 2.0-point spread on Originality exceeds all other dimensions.

Generative Self-Monitoring is null. All three attorneys score 5.0 on every pass in every condition. This is interpretable: Supreme Court oral argument is adversarial and time-constrained. Advocates cannot pause to self-correct. They prepare arguments in advance and execute them under pressure. The speech situation suppresses visible self-monitoring.

Epistemic Calibration inverts slightly. Dearing (6.3) marginally outscores both Clement and Shanmugam (6.0) on this dimension. A previous rubric version (with a different dimension architecture) scored Dearing highest overall when blinded — rewarding his systematic hedging and on-the-record qualifications. The reconceptualized Epistemic Calibration dimension in v3 scores *differentiated* confidence, not mere hedging, which corrected this inversion at the composite level while preserving it at the dimension level where it is informative rather than misleading.

Study 1: SCOTUS — Band Labels (Claude Blinded, First Pass)

DIMENSION	CLEMENT	SHANMUGAM	DEARING
Abstraction	Principled	Principled	Principled
Compression	Dense	Dense	Dense
Originality	Reframed	Inflected	Inflected
Conceptual Continuity	Cumulative	Cumulative	Cumulative
Epistemic Calibration	Differentiated	Differentiated	Differentiated
Gen. Self-Monitoring	Conceptual Revision	Conceptual Revision	Conceptual Revision

Five of six core dimensions produce identical band labels across all three speakers. Only Originality discriminates at the band level.

Study 1: SCOTUS — Reliability

DIMENSION	CLAUDE AVG SPREAD	GPT-4O AVG SPREAD
Abstraction	+/-0.3	+/-0.3
Compression	+/-0.3	+/-0.7
Originality	+/-0.3	+/-0.0
Conceptual Continuity	+/-0.3	+/-0.0
Epistemic Calibration	+/-0.3	+/-0.7
Gen. Self-Monitoring	+/-0.0	+/-0.0
Vocabulary	+/-0.0	+/-0.0
Syntax	+/-0.0	+/-0.0
Mean	+/-0.2	+/-0.2

Both models achieve high inter-pass reliability. Mean spread is +/-0.2 for both models. Claude's reliability improved substantially from earlier rubric versions (v1 showed +/-0.6 in some conditions), suggesting the behavioral descriptor matrix constrains scoring variability.

Study 2: MICASE — VRI Composite

CONDITION	FACULTY	GRADUATE	UNDERGRAD	PREDICTED ORDER?
Claude Unblinded	6.7 (108--116)	5.9 (103--111)	5.0 (96--104)	Yes
Claude Blinded	6.7 (108--116)	5.7 (101--109)	4.5 (92--100)	Yes

The predicted hierarchy holds in both conditions. The blinded VRI spread — 2.2 points from Faculty (6.7) to Undergraduate (4.5) — is the widest separation in any EC study to date, compared to 0.3 in the SCOTUS study. This confirms that EC discriminates across a wider ability range than within a narrow elite band.

Expectation effects. Faculty scores are identical in unblinded and blinded conditions (6.7/6.7). Graduate drops by 0.2 (5.9 to 5.7). Undergraduate drops by 0.5 (5.0 to 4.5). The pattern suggests the rubric gives modest benefit of the doubt to lower-tier speakers when their academic role is known — a rational Bayesian adjustment that disappears when the scorer operates blind.

Study 2: MICASE — Dimension-Level Scores (Claude Blinded, 3-Pass Average)

DIMENSION	FACULTY	GRADUATE	UNDERGRAD	SPREAD
Abstraction	7.0	6.7	6.0	1.0
Compression	6.0	4.0	3.7	2.3
Originality	8.0	5.7	5.0	3.0
Conceptual Continuity	7.0	6.0	5.0	2.0
Epistemic Calibration	6.0	6.3	4.0	2.3
Gen. Self-Monitoring	6.0	5.0	3.3	2.7
Vocabulary	7.0	7.0	5.0	2.0
Syntax	5.0	5.0	4.0	1.0

Every core dimension separates the three tiers. The widest spreads are on Originality (3.0), Generative Self-Monitoring (2.7), Compression (2.3), and Epistemic Calibration (2.3).

Generative Self-Monitoring: from null to strongest differentiator. In SCOTUS, all attorneys scored 5.0 on GSM with zero spread. In MICASE, Faculty scores 6.0, Graduate 5.0, Undergraduate 3.3 — a

2.7-point spread. The seminar context is exploratory and collaborative; speakers can pause, reconsider, and revise. The adversarial legal context suppresses this behavior. GSM's context-sensitivity is predicted by Levelt's self-monitoring model of speech production and constitutes evidence of construct validity: the dimension measures what it claims to measure, and it measures it only when the speech situation permits.

Epistemic Calibration ordering differs from VRI ordering. Graduate (6.3) slightly outscores Faculty (6.0) on this dimension even though Faculty's overall VRI is a full point higher. This is because the Graduate student's role in the seminar was to *critique* the Faculty member's theory — a task that inherently requires distinguishing stronger from weaker claims in someone else's argument. The Faculty member was *presenting* his own theory, which requires less visible epistemic differentiation. The tool is scoring the cognitive task being performed, not the speaker's credential.

Study 2: MICASE — Band Labels (Claude Blinded, First Pass)

DIMENSION	FACULTY	GRADUATE	UNDERGRAD
Abstraction	Principled	Principled	Relational
Compression	Dense	Sequential	Sequential
Originality	Generative	Reframed	Reframed
Conceptual Continuity	Cumulative	Connected	Connected
Epistemic Calibration	Differentiated	Differentiated	Occasionally Hedged
Gen. Self-Monitoring	Conceptual Revision	Conceptual Revision	Surface Repair

Faculty hits the highest band on 4 of 6 core dimensions. Graduate hits it on 2 of 6. Undergraduate hits it on 0 of 6. The band-label gradient provides a qualitative complement to the numerical scores.

Study 2: MICASE — Reliability

DIMENSION	AVG SPREAD (BLINDED)
Abstraction	+/-0.3
Compression	+/-0.3
Originality	+/-0.3
Conceptual Continuity	+/-0.0
Epistemic Calibration	+/-0.3
Gen. Self-Monitoring	+/-0.3

DIMENSION	AVG SPREAD (BLINDED)
Vocabulary	+/-0.0
Syntax	+/-0.0
Mean	+/-0.2

Reliability matches Study 1 exactly (mean +/-0.2). Per-pass VRI stability: Faculty 6.7/6.7/6.7, Graduate 5.7/5.4/5.9, Undergraduate 4.5/4.5/4.5. Faculty and Undergraduate produce identical VRI scores on all three passes; Graduate shows the most variation, consistent with its intermediate position on the ability continuum.

Cross-Study Comparison

PROPERTY	SCOTUS	MICASE
Speech context	Adversarial, time-pressured	Exploratory, collaborative
Ability range	Elite to competent	Faculty to undergraduate
VRI range (blinded)	5.9--6.2 (0.3 spread)	4.5--6.7 (2.2 spread)
Sharpest differentiator	Originality (2.0 spread)	Originality (3.0 spread)
Gen. Self-Monitoring	Null (all 5.0)	Strong (3.3--6.0, 2.7 spread)
Originality max	6 (Reframed)	8 (Generative)
Mean reliability	+/-0.2	+/-0.2
Predicted rank order (blinded)	Preserved	Preserved

DISCUSSION

The Single Claim

The EC scoring rubric detects known differences in verbal reasoning quality in the predicted direction, across two independent naturalistic speech corpora, at different points on the ability distribution. In Study 1, it correctly ranks three Supreme Court advocates by experience tier. In Study 2, it correctly ranks three academic speakers by academic level. Both studies use blinded scoring with high reliability. The construct generalizes beyond EC's own elicitation protocol.

Originality as the Sharpest Differentiator

Across both studies, Originality produces the widest dimension-level spread. In SCOTUS (blinded), Clement scores 6.0 while Dearing scores 4.0 — a 2.0-point gap. In MICASE (blinded), Faculty scores

8.0 while Undergraduate scores 5.0 — a 3.0-point gap. Originality is the only dimension in SCOTUS where the elite speaker receives a different band label from the other two even when the scorer does not know who it is reading.

This finding has theoretical significance. Originality, in the EC framework, is defined as "genuinely unexpected framing that illuminates — not obscurity, not mere novelty, but aptness combined with surprise." The highest band, Generative, requires the speaker to produce framings that "redefine the terms of the problem." The Faculty member in MICASE reaches this band by presenting original theoretical work on phenomenal illusions. Clement reaches "Reframed" by finding novel angles in legal argument that the Court has not previously considered.

What makes this finding important for construct theory is that Originality is not a proxy for general ability. It responds to the cognitive task being performed, not the speaker's credential. If a lower-tier speaker were performing a more generative cognitive task than a higher-tier speaker, we would expect Originality to invert the expected hierarchy — and in an earlier version of the MICASE study (Study 2a, not reported here in detail), this is precisely what occurred. An undergraduate performing novel literary analysis scored higher on Originality than a graduate student performing routine methodological critique. The VRI composite still placed the graduate student higher, but the Originality dimension captured something real about the cognitive task the undergraduate was performing.

This demonstrates that EC is not a g-proxy. It does not simply assign higher scores to higher-status speakers. Dimension profiles carry independent information about *what kind of reasoning* is happening, and that information can diverge from the composite in theoretically motivated ways.

Generative Self-Monitoring: Context-Sensitivity as Construct Validity

GSM's behavior across the two studies is the second theoretically significant finding. In SCOTUS, all three attorneys receive identical GSM scores (5.0, zero spread) across all conditions and both models. In MICASE, GSM produces the second-widest spread of any dimension (2.7 points) and cleanly separates all three academic tiers.

The explanation is straightforward: Supreme Court oral argument does not permit visible self-correction. Advocates prepare their arguments, execute them under time pressure, and cannot pause to revise. The speech situation suppresses the behavior GSM is designed to detect. Academic seminars, by contrast, are exploratory. Speakers can pause, reconsider, qualify, and revise upward. The behavior is permitted, and GSM detects it.

This context-sensitivity is not a flaw in the dimension. It is predicted by Levelt's (1983) self-monitoring model of speech production, which holds that speakers monitor their own output and initiate repairs when monitoring detects errors or suboptimal formulations. The rate and quality of

self-monitoring varies with the speech situation: it is highest in low-stakes exploratory contexts and lowest in high-stakes performative contexts.

A dimension that produced identical scores regardless of context would be measuring something other than self-monitoring. GSM's sensitivity to context is evidence that it measures what it claims to measure.

Epistemic Calibration: Scoring the Task, Not the Title

In MICASE, the Graduate student (6.3) slightly outscores the Faculty member (6.0) on Epistemic Calibration even though the Faculty member's overall VRI is a full point higher. This is because the Graduate student's task — critiquing someone else's theory — inherently requires distinguishing stronger from weaker claims, identifying assumptions, and marking the boundary between established knowledge and speculation. The Faculty member's task — presenting his own theory — calls for confidence and cumulative argument-building, not epistemic differentiation.

The rubric is correct to score the Graduate student higher on this dimension. It is detecting the cognitive demands of the task being performed, not the credential of the person performing it. This result, combined with the Originality findings, establishes that EC scores verbal reasoning *as performed in a specific context*, not speaker ability in the abstract.

Why These Two Studies Belong Together

SCOTUS and MICASE make the same argument at different ends of the ability spectrum. SCOTUS tests whether EC detects fine-grained differences among elite speakers operating within a narrow band. MICASE tests whether EC detects coarse-grained differences across the full academic range. The affirmative answer in both cases establishes that EC's validity is not range-restricted.

The two corpora also provide complementary evidence about the dimension structure. SCOTUS reveals which dimensions compress in adversarial, performative contexts (GSM goes to zero; most dimensions converge). MICASE reveals which dimensions expand in exploratory, collaborative contexts (GSM becomes the strongest differentiator; all dimensions spread). Together, they map the conditions under which each dimension is maximally and minimally informative.

LIMITATIONS

- **Small N per tier.** Each study has only one speaker per tier. The findings are consistent and reliable (3 passes, high inter-pass stability), but they represent three individuals, not three samples from a population. Generalization to "all elite advocates" or "all faculty" requires replication with larger samples.
- **Task confound in MICASE.** Faculty and Graduate were in the same seminar but performing different cognitive tasks (presenting vs. critiquing). The Undergraduate was in a different seminar

entirely. We interpret this as informative — the rubric scores the reasoning being performed, not the speaker's title — but a design with all three speakers performing the same task would provide cleaner evidence.

- **Single-model scoring in MICASE.** Study 1 uses both Claude and GPT-4o; Study 2 uses only Claude. Cross-model replication in MICASE would strengthen confidence, though Claude's superior blinded performance in Study 1 motivated its selection as the sole scorer.
- **No external criterion variable.** These studies test whether EC detects known-group differences, not whether EC scores predict any outcome. Predictive validity and ecological validity are addressed in separate work (the CWT Study, forthcoming as Paper 3).
- **Transcript length varies.** Dearing's SCOTUS transcript (3,886 words) is 3--4x longer than Clement's (1,141 words) or Shanmugam's (950 words). The Undergraduate MICASE transcript (746 words) is less than half the Faculty member's (1,973 words). Length differences could affect scores, though the rubric instructs the scorer to evaluate sustained performance, not quantity of output.
- **LLM scoring is not human scoring.** The rubric is applied by Claude Sonnet 4 and GPT-4o, not by trained human raters. LLM scoring introduces its own sources of error that differ from human rater variability. The models may share systematic biases that human raters would not exhibit. However, the blinding protocol and multi-pass design mitigate the most obvious such bias (name recognition), and the high reliability suggests the descriptor matrix constrains the models effectively.

CONCLUSION

Across two naturalistic speech corpora — Supreme Court oral arguments and university seminar discussions — the EC scoring rubric detects known differences in verbal reasoning quality in the predicted direction, with high reliability and minimal expectation effects. The construct generalizes beyond EC's own elicitation protocol.

Two emergent findings contribute to construct theory. First, Generative Self-Monitoring is context-sensitive in a way predicted by Levelt's speech production model: null in adversarial argument, strongly differentiating in exploratory discussion. This is evidence that the dimension measures what it claims to measure. Second, Originality and Epistemic Calibration can diverge from the overall VRI hierarchy when the cognitive task performed by a lower-tier speaker is more demanding on those specific dimensions. This establishes that EC scores verbal reasoning as performed, not speaker credential, and that dimension profiles carry independent information beyond the composite.

The EC rubric is not a g-proxy. It is a multi-dimensional instrument that detects both the level and the kind of verbal reasoning occurring in spontaneous speech.

Study 1 data: 3 SCOTUS advocates, 36 scoring passes (2 models x 2 conditions x 3 passes x 3 speakers) **Study 2 data:** 3 MICASE speakers, 18 scoring passes (1 model x 2 conditions x 3 passes x 3 speakers) **Models:** Claude Sonnet 4 (claude-sonnet-4-20250514), GPT-4o (OpenAI) **Rubric version:** v3 (reconceptualized 6+2 dimension architecture with behavioral descriptor matrix)