SCOTUS Scaled Validity Study: Construct Generalizability in High-Stakes Professional Discourse (n=12)

ABSTRACT

This study scales the original SCOTUS construct validity test from 3 to 12 advocates across three experience tiers (elite, experienced, junior), scored by two independent LLMs (Claude Sonnet 4 and GPT-4o) in blinded conditions. It adds a written-vs-spoken comparison using legal briefs from the same cases as the oral arguments, and correlates VRI scores with a biographical proxy for cognitive ability. The rubric correctly ranks tier averages on Claude (elite 6.3 > junior 6.1 > experienced 6.0) with VRI correlating r=0.688 with the biographical proxy. The written-vs-spoken comparison reveals that SCOTUS advocates' EC profiles are largely stable across modalities — unlike the CWT sample, where tech founders' profiles collapsed in writing. Originality remains the sharpest dimension-level differentiator, and GSM remains context-suppressed in adversarial argument.

METHOD

Sample

12 advocates extracted from Oyez oral argument transcripts, plus 3 from the original v3 study (Clement, Shanmugam, Dearing), for a total of 15 transcripts from 13 unique advocates.

Elite tier (30+ SCOTUS arguments):

ADVOCATE	ARGUMENTS	CASE SCORED	WORDS
Paul Clement	~92	NetChoice v. Paxton (2024)	1,141
Neal Kumar Katyal	~45	Trump v. Hawaii (2018)	3,926
Neal Kumar Katyal	—	Moore v. Harper (2022)	5,084
Seth P. Waxman	~80	SFFA v. Harvard (2022)	5,585
Seth P. Waxman	—	Samsung v. Apple (2016)	2,683
Theodore B. Olson	~65	Hollingsworth v. Perry (2013)	2,284

Experienced tier (5-20 arguments):

ADVOCATE	CASE SCORED	WORDS
Kannon Shanmugam	Seila Law v. CFPB (2020)	950
Scott G. Stewart	Dobbs v. Jackson (2021)	4,482
Julie Rikelman	Dobbs v. Jackson (2021)	2,998
Kristen K. Waggoner	303 Creative v. Elenis (2022)	4,361
Damien M. Schiff	Sackett v. EPA (2022)	5,609

Junior tier (1-3 arguments):

ADVOCATE	CASE SCORED	WORDS
Richard Dearing	NY Rifle & Pistol v. NYC (2019)	3,886
Eric R. Olson	303 Creative v. Elenis (2022)	3,940
Cameron T. Norris	SFFA v. Harvard (2022)	3,922
David H. Thompson	Moore v. Harper (2022)	5,624

Biographical Proxy for Cognitive Ability

Each advocate was assigned a proxy score (1-10) based on publicly available biographical data, combining law school tier, Supreme Court clerkship, law review membership, and career achievement markers. This composite serves as a rough proxy for fluid/crystallized intelligence — law school admission correlates ~0.5 with LSAT, which correlates ~0.5 with IQ.

PROXY SCORE	ADVOCATES
10	Clement, Katyal, Waxman, Shanmugam
9	Olson, Norris
8	Rikelman
7	Dearing, Thompson
6	Stewart
5	Waggoner, Schiff, E. Olson

Scoring Protocol

Identical to the original v3 study. Claude Sonnet 4 and GPT-4o each scored all transcripts blinded (randomized speaker labels, no identifying information), 3 passes per advocate, with 12-second delays between calls.

Written-vs-Spoken Comparison

For 4 advocates (Katyal, Waxman, Olson, Stewart), legal briefs from the same case as the oral argument were extracted, trimmed to ~2,000 words, and scored blinded as "Author X" with no identifying information. This provides a same-person, same-case, same-legal-issues comparison where the only variable is the modality (written brief vs. spoken argument).

RESULTS

Tier Separation

Claude Sonnet 4 — Blinded, Unique Advocates (n=10):

ADVOCATE	TIER	PROXY	ABSTR	COMPR	ORIG	CONT	EPIST	GSM	VOCAB	SYNTAX
Katyal	Elite	10	7	6.9	6	7.9	6	5	7.9	6.9
Norris	Junior	9	7	6.3	6	7.3	6	5	7.3	6.3

ADVOCATE	TIER	PROXY	ABSTR	COMPR	ORIG	CONT	EPIST	GSM	VOCAB	SYNTAX
Olson	Elite	9	7	6	6	7	6	5	7.3	6.3
Schiff	Experienced	5	7	6	5.3	7	6	5	7.3	6.3
Thompson	Junior	7	7	6	5.7	7	6	5	7.3	6.3
Waxman	Elite	10	7	6	5	7	6	5.5	7.4	6.4
Rikelman	Experienced	8	7	6	4.7	7	6	5	7	6
Waggoner	Experienced	5	7	6	5	7	6	5	7	6
E. Olson	Junior	5	7	6	5	7	6	5	7	6
Stewart	Experienced	6	7	6	4.7	7	6	4.3	7	6

Tier Averages (Claude, unique advocates):

TIER	N	VRI MEAN	PROXY MEAN
Elite	3	6.3	9.7
Junior	3	6.1	7.0
Experienced	4	6.0	6.0

The elite tier averages highest on VRI, but the spread is narrow (6.0--6.5). The experienced tier averages lowest, not the junior tier — this reflects the fact that some junior advocates (Norris) have elite credentials (Yale Law, SCOTUS clerkship) despite few arguments, while some experienced advocates (Waggoner, Schiff) have lower-tier credentials.

VRI Correlates with Biographical Proxy

MEASURE	R (N=10)
VRI vs. Proxy	**0.688**
Compression vs. Proxy	0.526
Originality vs. Proxy	0.504
Continuity vs. Proxy	0.526
Vocabulary vs. Proxy	0.659
Syntax vs. Proxy	0.659
GSM vs. Proxy	0.429
Abstraction vs. Proxy	0.000
Epistemic Cal vs. Proxy	0.000

VRI correlates r=0.688 with the biographical proxy — a strong positive correlation. This is substantially higher than the CWT study's r=0.298, likely because the SCOTUS proxy is a more direct measure of the same underlying ability (legal reasoning quality) than the CWT reputation score (general intellectual status).

Abstraction and Epistemic Calibration show zero correlation with proxy. Abstraction ceilings at 7 for every advocate (same pattern as in CWT). Epistemic Calibration is flat at 6 for everyone — adversarial argument compresses this dimension, as established in the original v3 study.

Vocabulary and Syntax correlate highest with proxy (both r=0.659). This is expected in the legal domain — more credentialed advocates use more precise legal vocabulary and more complex sentence structures. In the CWT study, these moderators showed near-zero correlation with reputation, suggesting the SCOTUS proxy captures something more education/training-correlated than the CWT reputation score.

Cross-Model Comparison

GPT-4o tier averages (blinded):

TIER	GPT-4O VRI	CLAUDE VRI	DELTA
Elite	6.5	6.3	+0.2
Experienced	6.5	6.0	+0.5
Junior	6.7	6.1	+0.6

GPT-4o scores higher overall and does not separate tiers. Its junior tier average (6.7) exceeds its elite tier average (6.5). Claude correctly separates tiers; GPT-4o does not. This replicates the original v3 finding that Claude is the more discriminating scorer in blinded conditions.

Written vs. Spoken (n=4)

DIMENSION	KATYAL W/S	WAXMAN W/S	OLSON W/S	STEWART W/S	AVG DELTA
Abstraction	8/8 (0)	7/7 (0)	7/8 (-1)	8/7 (+1)	**0.0**
Compression	7/7 (0)	7/6 (+1)	6/7 (-1)	8/6 (+2)	+0.5
Originality	6/7 (-1)	4/5 (-1)	4/6 (-2)	6/5 (+1)	**-0.8**
Continuity	8/8 (0)	8/7 (+1)	7/8 (-1)	8/7 (+1)	+0.3
Epistemic Cal	7/7 (0)	7/6 (+1)	6/7 (-1)	7/6 (+1)	+0.3
GSM	6/6 (0)	6/6 (0)	5/7 (-2)	7/5 (+2)	**0.0**
Vocabulary	8/8 (0)	8/7 (+1)	7/8 (-1)	8/7 (+1)	+0.3
Syntax	8/7 (+1)	8/6 (+2)	7/7 (0)	8/6 (+2)	**+1.3**

Two dimensions are stable across modalities: Abstraction (avg delta 0.0) and GSM (avg delta 0.0). These measure the person, not the medium.

Syntax increases in writing (+1.3 avg). Legal briefs are syntactically more complex than oral argument. This is expected — briefs undergo multiple rounds of editing for sentence-level precision.

Originality decreases in writing (-0.8 avg). Briefs are more formulaic than oral argument. Advocates follow established briefing conventions; at oral argument, they must respond to unexpected questions from the justices, which elicits more original framing.

The Olson anomaly. Olson scores substantially higher on spoken (VRI 7.2) than written (VRI 5.9) — a 1.3-point gap. His GSM drops from 7 in speech to 5 in writing. This suggests Olson is an especially strong oral advocate whose spontaneous reasoning exceeds his written product. This is the inverse of the CWT finding for Collison (whose writing exceeded his speech). The two cases together demonstrate that the modality gap is person-specific, not a systematic bias.

Katyal is perfectly stable. Six of eight dimensions are identical across modalities. His brief and his oral argument produce the same cognitive fingerprint. This is convergent validity evidence — the same construct measured two different ways produces the same result for this advocate.

DISCUSSION

What the Scaled Study Adds

The original SCOTUS v3 study (n=3) showed the rubric could rank three advocates in the predicted order. The scaled study (n=10 unique advocates) shows three additional things:

- **VRI correlates strongly with a biographical cognitive ability proxy (r=0.688).** This is the first evidence linking EC scores to an external measure of ability in the SCOTUS domain. The correlation is driven by Compression, Originality, Continuity, and Vocabulary — not by Abstraction (which ceilings) or Epistemic Calibration (which is context-suppressed).
- **Tier assignment by argument count is not the best predictor.** The proxy score (which combines credentials with experience) outpredicts the tier label. Some junior advocates with elite credentials (Norris: Yale Law, SCOTUS clerkship, VRI 6.3) score above experienced advocates with lower credentials (Stewart: VRI 5.9). The rubric is scoring the person's reasoning quality, not their argument count.
- **Claude discriminates; GPT-4o does not.** At n=10, Claude correctly ranks tiers (elite > junior > experienced, driven by credential effects). GPT-4o produces flat or inverted tier averages. This confirms Claude as the appropriate scorer for blinded construct validity studies and justifies its selection as EC's production scorer.

Written-vs-Spoken: SCOTUS Advocates vs. CWT Guests

The CWT written-vs-spoken pilot (n=3) showed that Andreessen's GSM and Epistemic Calibration collapsed from speech to writing (-2 each). His manifesto was declarative; his conversational speech preserved more self-correction and hedging.

The SCOTUS written-vs-spoken comparison (n=4) shows a different pattern: **on average, SCOTUS advocates' profiles are stable across modalities.** Abstraction and GSM both average 0.0 delta. The one consistent shift is Syntax (+1.3 in writing), which reflects the editorial polish of legal briefs.

Why the difference? SCOTUS advocates are trained to maintain the same reasoning style in both media. Their briefs and their oral arguments serve the same adversarial purpose and make the same legal claims. CWT guests, by contrast, produce writing (blogs, manifestos, academic papers) in a fundamentally different mode from their conversational speech. The modality gap is larger when the two contexts serve different communicative purposes.

This finding has implications for EC's assessment design: the verbal assessment measures reasoning-as-performed-in-conversation. For people whose conversational and written reasoning styles

converge (Katyal, professional advocates), EC scores generalize to written contexts. For people whose styles diverge (Andreessen, tech founders), EC specifically measures the conversational mode — which is the mode where GSM and Epistemic Calibration are most visible.

Limitations

- **The proxy is not IQ.** The biographical proxy correlates with LSAT (which correlates with IQ), but it is not a direct measure of fluid intelligence. The Gf correlation study (using Raven's or WAIS) remains necessary.
- **Small n per tier.** 3-4 advocates per tier is an improvement over n=1, but still small. Effect sizes should be interpreted cautiously.
- **Selection of written samples.** The Olson brief came from a different case than his oral argument (Perry answering brief vs. Hollingsworth oral argument — same underlying case but different procedural posture). This may explain some of his anomalous written-spoken gap.
- **GSM remains context-suppressed.** All advocates score 4.3-5.5 on GSM in oral argument, with minimal variation. This dimension cannot discriminate in adversarial contexts, as established in the original v3 study.

CONCLUSION

At n=10 unique advocates scored by two models, the SCOTUS scaled study confirms and extends the original v3 findings. VRI correlates r=0.688 with a biographical cognitive ability proxy. Claude correctly separates tiers while GPT-4o does not. The written-vs-spoken comparison shows that SCOTUS advocates' profiles are largely stable across modalities, providing convergent validity evidence for the rubric.

The key limitation of the SCOTUS context remains: adversarial argument suppresses GSM and compresses Epistemic Calibration, leaving Originality as the only core dimension with meaningful individual variation. This makes the SCOTUS corpus ideal for testing construct generalizability (does the rubric detect anything at all in this context?) but less ideal for studying the full dimensional structure. The MICASE and CWT corpora, with their exploratory and conversational contexts, provide richer data for understanding how the six dimensions interact.

Transcripts: Oyez Project (oyez.org), Supreme Court oral argument audio transcripts **Briefs:** Supreme Court of the United States, filed merits briefs **Models:** Claude Sonnet 4 (claude-sonnet-4-20250514), GPT-4o (OpenAI) **Rubric version:** v3 (6 core + 2 moderator dimensions) **Total scoring passes:** 72 (oral arguments) + 8 (briefs) = 80