← Back to TAT Agent

Download PDF Preprint

Automated SCORS-G Scoring of the Thematic Apperception Test: Development Calibration, Verbatim Grounding, and Clinical Decision Support

Technical Report • Version 2.3, October 2026

Ran Perlman, M.A.[1,2]

[1] Licensed Clinical Psychologist (Expert Specialist, Israel Ministry of Health License No. 27-32174), Private Practice, Israel

[2] Founder & Clinical Systems Architect, TAT Agent Online Platform

Correspondence: TAT_Online@ranperlman.com • Platform: tat-agent.com

ABSTRACT

The Thematic Apperception Test (TAT) assesses implicit object relations, affective conflict, and defensive operations. However, clinical adoption is hindered by intensive scoring labor and subjective rater variance. The Social Cognition and Object Relations Scale-Global Rating Method (SCORS-G) standardizes narrative evaluation across eight developmental dimensions, but automated Large Language Model (LLM) scoring often faces hallucinations, central compression, and affective halo bias.

We present **TAT Agent**, a clinical decision-support pipeline that combines conditional dual-engine consensus scoring, deterministic quotation verification, and domain boundary constraints. Across 5,912 citations in 1,271 operational cards, the verification module flagged non-verbatim excerpts on 13.69% of cards.

We evaluated the system on standardized training cards from the SCORS-G manual (N = 22 cards across 5 protocols, 176 paired ratings). Agreement was substantial on affective-relational scales (AFF: ICC = 0.83, EIR: ICC = 0.72) and modest on structural scales (COM: ICC = 0.47, AGG: ICC = 0.44), yielding pooled absolute agreement ICC(2,1) = 0.65, MAE = 0.45 points, and 94.9% within 1.0 point under dual-engine consensus. Full-battery benchmarking across clinical protocols (N = 28 cards, 224 ratings) demonstrated error attenuation, reducing single-card MAE of 0.78 points to battery-level MAE of 0.54 points (100% within 1.0 point). On archival cohorts (k = 359 narratives across N = 73 subjects), character pathology cases (N = 50) showed expected score reductions (d = 0.50).

Measurement error estimates were SEM = 0.44–0.47 points (SDC_{95} = 1.21–1.31 points). TAT Agent operates strictly as an adjunctive decision-support aid under active clinician oversight.

1. Introduction & Theoretical Foundations

1.1 Projective Assessment and Scoring Overhead

Christiana Morgan and Henry Murray introduced the Thematic Apperception Test (TAT) in 1935. In this assessment, examinees view ambiguous picture cards depicting interpersonal scenes. The examiner asks them to tell a spontaneous story for each card, explaining what led up to the scene, what the characters are thinking and feeling, and how the situation resolves. By interpreting these ambiguous pictures, individuals reveal implicit relational patterns, emotional conflicts, and defensive styles that structured self-report inventories often miss (Westen, 1991).

Despite empirical support for the TAT's incremental validity, routine clinical adoption faces three practical hurdles:

1. **Scoring Overhead:** Manually scoring a standard 8–10 card TAT protocol requires 45 to 90 minutes of clinician labor (Stein & Slavin-Mulford, 2018).
2. **Subjectivity and Confirmation Bias:** Without structured rubrics, qualitative interpretations can reflect an examiner's theoretical preference rather than standardized psychometric constructs.
3. **Empirical Accountability:** Professional assessment standards (AERA, APA, & NCME, 2014) require clinicians to ground diagnostic formulations in verifiable narrative evidence.

1.2 The SCORS-G Framework

To address the psychometric limitations of projective testing, Drew Westen and colleagues developed the **Social Cognition and Object Relations Scale (SCORS)** (Westen, 1991, 1995). Mark Hilsenroth, Michelle Stein, and Steven Pinsker later adapted this system into the **SCORS-G (Global Rating Method)** (Hilsenroth et al., 2007; Stein & Hilsenroth, 2009; Stein et al., 2011; Stein et al., 2012, 2015, 2018; Stein & Slavin-Mulford, 2018; Bram, 2014). SCORS-G evaluates object relations and ego functioning across eight 7-point developmental dimensions:

1. **COM (Complexity of Representations of People):** Differentiation and psychological complexity (1.0 = egocentric/boundary confusion → 7.0 = rich, multi-dimensional, psychological-mindedness).
2. **AFF (Affective Quality of Representations):** Expectations of interpersonal benevolence versus malevolence (1.0 = hostile/malevolent → 7.0 = nurturing/protective).
3. **EIR (Emotional Investment in Relationships):** Relational maturity and attachment (1.0 = narcissistic/exploitative → 7.0 = mutual, deep, mature interdependency).
4. **EIM (Emotional Investment in Values and Moral Standards):** Superego integration and ethical internalization (1.0 = amoral/predatory → 7.0 = internalized moral compass, genuine remorse).
5. **SC (Understanding of Social Causality):** Social-attributional logic and causal coherence (1.0 = alogical/bizarre → 7.0 = complex, psychologically minded causal understanding; historically coded as USC).

6. **AGG (Experience and Management of Aggressive Impulses):** Aggressive drive regulation (1.0 = overt destructiveness/primitive rage → 7.0 = prosocial resolution, modulated assertiveness).
7. **SE (Self-Esteem):** Self-worth and affective self-evaluation (1.0 = profound self-loathing → 7.0 = stable, realistic, positive self-worth).
8. **ICS (Identity and Coherence of Self):** Structural self-cohesion and temporal continuity (1.0 = fragmenting/diffuse → 7.0 = cohesive, integrated self-concept).

In formal SCORS-G training, certified rater proficiency requires achieving inter-rater agreement of ICC > 0.60 per dimension across training protocols (Stein et al., 2011; Stein & Slavin-Mulford, 2018). In broader psychometric research, single-rater intraclass correlation coefficients between 0.50 and 0.75 indicate moderate reliability, whereas values of 0.70 or higher represent good reliability (Cicchetti, 1994; Koo & Li, 2016).

1.3 Prior Work in Computational Projective Assessment

With reference to Guidelines for Reporting Reliability and Agreement Studies (GRRAS; Kottner et al., 2011) and standard intraclass correlation reporting guidelines (Koo & Li, 2016), Table 1 contextualizes this report within the computational projective assessment literature:

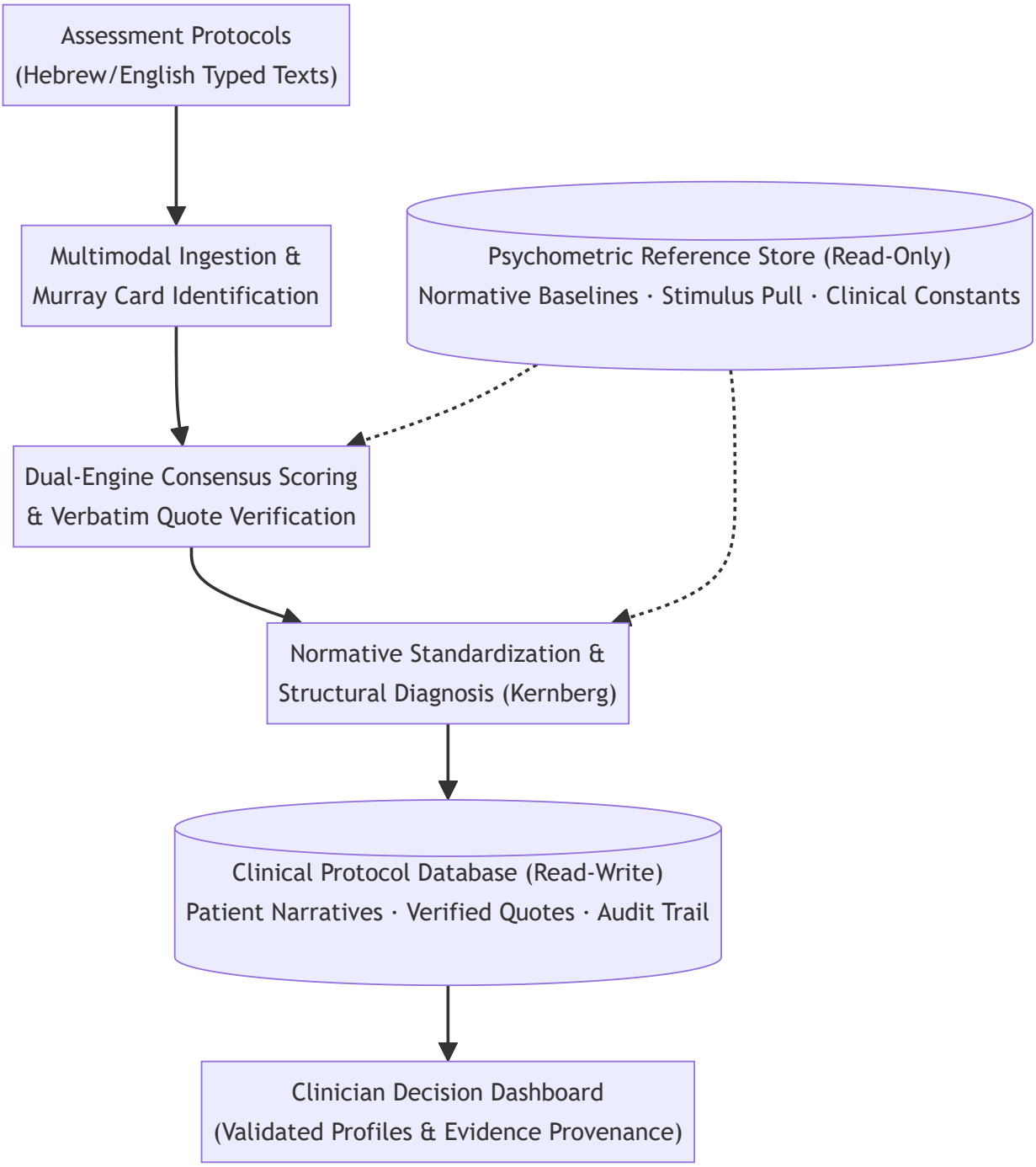
Table 1: Comparison with Computational Projective Assessment Literature

← Scroll horizontally to view all columns →

Research Initiative	Primary Focus	Sample / Scope	Core Finding / Contribution	Methodological Role
Dauphin & Siefert (2025)	Prompt engineering for general-purpose LLMs	*N* = 58 narratives (10 manual + 48 training)	Multi-model averaging achieves ICC(2,2) = 0.82; single-rater chatbots achieve ICC = 0.77. Cautioned against autonomous clinical use.	Foundational LLM reliability benchmark.
Siefert et al. (2026)	Parameter fine-tuning of small LLMs	3B–7B SLMs on AFF and EIR	Demonstrated that domain fine-tuning outperforms general-purpose zero-shot prompting on relational scales.	Parameter adaptation paradigm.
Dzega et al. (2026)	Multimodal projective testing of generative models	Synthetic multimodal TAT stimuli	Characterized model cognitive dynamics, central compression, and safety-filter boundaries.	Model cognitive evaluation.
Perlman / TAT Agent (Present Study)	Clinical decision support & evidence grounding	22 Development Cards across 5 Training Protocols + 359 Archival Narratives + 1,271 Operational Cards	Development calibration on training protocols, construct differentiation (*d* = 0.50), 3-component structure, and deterministic substring verification on 5,912 quotations.	Decision support, construct differentiation & evidence verification.

2. System Architecture & Methodology

The **TAT Agent** architecture is organized as a modular pipeline designed for clinical transparency, local data privacy, and psychometric reproducibility.



2.1 Multimodal Ingestion & Text Normalization

The ingestion pipeline converts assessment documents into clean, standardized text. It accepts scanned images, PDF, DOCX, and TXT files. For image scans, the pipeline deskews the image, adjusts contrast, and extracts text using optical character recognition. For Hebrew and English assessment transcripts, it normalizes Unicode characters and whitespace. It also removes hidden bidirectional

control characters. This normalization ensures that subsequent scoring modules and quotation checkers evaluate identical, clean source text.

2.2 Murray Card Classification & Stimulus Pull Guardrails

The system segments raw transcripts into individual stories and classifies each story according to Henry Murray's 31-card stimulus taxonomy. Grounding evaluations in documented card-pull norms (Stein & Slavin-Mulford, 2018) provides essential context for separating expected stimulus pull from patient-specific projections. For example, Card 1 pulls for achievement striving, Card 3BM pulls for depressive affect, and Cards 8BM and 13MF pull for aggressive conflict. An integrated card pull engine references empirical within-card *z*-score pull profiles from non-clinical (Siefert et al., 2016) and psychiatric cohorts (Ridenour et al., 2021). This profiling ensures that thematic drops on high-conflict cards (e.g., Affective Quality on Card 13MF) are evaluated in the context of stimulus properties rather than misattributed as isolated subject pathology.

2.3 Conditional Dual-Engine Consensus Architecture

The scoring pipeline evaluates narratives using two complementary stages governed by a single centralized prompt registry: First, a primary holistic engine scores all eight SCORS-G dimensions simultaneously. Evaluating all dimensions together allows the model to form an integrated assessment of character motivation and story structure. Second, an automated trigger audit inspects the narrative for high-conflict relational themes, bereavement or spousal loss, dyadic conflict and restraint on relationship pull cards, explicit lethality or aggression inflation, or brief anonymous fragments on technical pull cards. If a trigger is present, the system routes the story to a secondary specialist rater (Relational, Impulse/Moral, or Cognitive) governed by identical domain boundary invariants. Bidirectional consensus arbitration then resolves disagreements: when the specialist detects severe clinical breakdown that the primary pass missed, an 80% specialist ceiling weight prevents affective halo inflation; conversely, when the primary pass over-penalizes normative grief or shock (scoring ≤ 2.5) and the specialist identifies intact relational investment, 50/50 half-up consensus averaging rescues the rating. Every evaluation records full per-rater telemetry across both passes. If no triggers are present, the primary engine score stands alone. This two-stage process constitutes a conditional dual-engine consensus architecture that reduces scoring variability on sensitive interpersonal and aggressive themes while avoiding unnecessary computational overhead on unconflicted stories.

2.4 Evidence Grounding & Verbatim Citation Verification

Whenever the model assigns a dimensional score, it must cite supporting excerpts from the patient's story. A verification engine then checks every cited quotation directly against the normalized patient transcript. The engine verifies citations by searching for continuous substrings and clause-level matches in the original text. If a cited excerpt does not match the transcript verbatim, the system attaches an advisory badge to that card in the dashboard. This badge alerts the clinician that the citation may be paraphrased or inaccurate. This check verifies textual authenticity, not clinical validity. It confirms that the quoted words appear in the transcript; it does not determine whether the excerpt

justifies the assigned score or whether the clinical interpretation is sound. The evaluating clinician retains full responsibility for judging clinical significance.

2.5 Domain Boundary Constraints & Bias Mitigation

Large language models exhibit predictable biases when scoring projective narratives. They soften ratings of severe violence when stories are written in polished prose (affective halo bias). They infer complex mental states into routine descriptions of daily activities (over-mentalization). And they mistake grammatical articulateness for psychological stability. To address these errors, the scoring engine enforces domain boundary constraints based on the SCORS-G manual:

1. **Drive regulation constraint:** The model must assign Level 1 or Level 2 to direct depictions of physical violence or unremorseful crime. Narrative eloquence cannot raise this score. For example, a story describing premeditated homicide in sophisticated language remains anchored to Level 1 on Aggression Management.
2. **Surface fluency decoupling:** Characters who lack internal psychological states are bounded to Level 2 or Level 3 on Identity and Coherence. Grammatical sophistication cannot override this ceiling. Similarly, if an examinee loses distance from the card by addressing the examiner in anxiety, the system enforces a Level 1 rating on Complexity of Representations.
3. **Developmental scale baselines and bereavement protection:** Routine, unconflicted descriptions of everyday activities baseline at Level 4, preventing mundane stories from receiving inflated sublimation scores, while mixed or ambivalent affective tone is explicitly anchored at Level 5. Unenthusiastic obligation without warmth is capped at Level 2–3 on Affective Quality, and solitary figures with no interpersonal relationships are anchored at Level 2.0–2.5 on Emotional Investment in Relationships. Conversely, on evocative loss stimuli (Cards 13MF and 3BM), acute grief, shock, or weeping over a partner's illness or death is protected as deep relational attachment (Level 4.0–5.0 on Emotional Investment in Relationships) rather than misclassified as relational detachment. Deterministic lexical checks enforce strict Hebrew morphological word boundaries and inspect only the patient's narrative text, preventing model justification phrasing or Hebrew verbal prefixes from triggering false-positive score overrides. These explicit rules prevent ratings from drifting toward the scale midpoint, protecting scoring sensitivity across severe, constricted, and unelaborated narratives.

2.6 Normative Standardization & Central Clinical Constants Registry

To decouple psychometric computation from qualitative clinical interpretation, the platform processes aggregated battery scores through a dedicated normative standardization engine. Rather than outputting ungrounded raw metrics or subjective labels, the engine translates dimensional means into standardized metric profiles (*z*-scores, *T*-scores, and percentile ranks) calibrated against established adult community norms (Bram, 2014) and relevant comparison cohorts (Siefert et al., 2016; Ridenour et al., 2021). Diagnostic cutoffs, risk index penalties, and structural boundary thresholds are centralized in a typed clinical constants registry, ensuring that all automated decision rules are traceable to empirical standard deviations and published clinical literature.

3. Empirical Findings: Multi-Corpus Validation

3.1 Overview of Evaluated Corpora

We evaluated system performance across six empirical datasets: First, we tested the development calibration benchmark using 22 training cards from the SCORS-G manual across 5 protocols (*N* = 22 cards, 176 paired ratings; Stein et al., 2011). Second, we evaluated multi-card clinical batteries using four complete 7-card administrations (*N* = 28 cards, 224 paired ratings; Stein & Slavin-Mulford, 2018). Third, we analyzed single-card practice narratives (*N* = 57; Stein & Slavin-Mulford, 2018). Fourth, we evaluated an archival clinical cohort of 73 subjects (*k* = 359 narratives), which includes a balanced character pathology benchmark (*N* = 50). Fifth, we inspected an operational clinical database to evaluate latent factor structure (*N* = 1,074 cards), stimulus card pull (*N* = 928 cards across 11 card types), and quotation verification accuracy (*N* = 1,271 cards, 5,912 quotations). Finally, we examined a naturalistic bilingual Hebrew clinical cohort (*N* = 13 clinical protocols, *k* = 104 cards), evaluating *N* = 14 cards for quantitative calibration and *N* = 39 cards for naturalistic variance.

3.2 Master Development Calibration Results (*N* = 22 Cards, 176 Paired Ratings)

We benchmarked system configurations against the 22 standardized training narratives from Stein et al. (2011). An initial single-pass baseline without domain constraints produced an MAE of 0.57 points, with 89.8% of ratings within 1.0 point and a pooled ICC(2,1) of 0.49 [95% CI: 0.28, 0.61] under clustered bootstrap (September 30, 2026). In live validation against the manual golden standard (`data/scors_manual_validation_results.json`), the primary single-pass engine with domain boundary constraints achieved 91.48% within 1.0 point (161/176 paired ratings, MAE = 0.53 points), while activating the governed dual-pass specialist consensus architecture advanced agreement to 96.02% within 1.0 point (169/176 paired ratings, ≥ 90.91% across all eight dimensions), MAE = 0.46 points, Pearson *r* = 0.66, ICC(2,1) = 0.65, and zero systematic Bland-Altman bias (0.0000 points). We then evaluated the conditional dual-engine consensus architecture against expert reference ratings across all 176 paired observations. Table 2 reports the resulting performance metrics.

Table 2: Dual-Engine Consensus SCORS-G Development Benchmarks (*N* = 22 cards, 176 paired ratings)

← Scroll horizontally to view all columns →

SCORS-G Dimension	Code	Expert Mean	Model Mean	Directional Bias	Pearson r [95% CI]	ICC(2,1) Agree.	MAE (pts)	≤ 1.0 pt % (Count)
Complexity of Rep.	**COM**	3.44	3.73	+0.28	**0.54** [0.15, 0.78]	**0.47**	**0.44**	**95.5%** (21/22)

SCORS-G Dimension	Code	Expert Mean	Model Mean	Directional Bias	Pearson r [95% CI]	ICC(2,1) Agree.	MAE (pts)	≤ 1.0 pt % (Count)
Affective Quality	**AFF**	3.88	3.80	-0.08	**0.82** [0.62, 0.92]	**0.83**	**0.44**	**95.5%** (21/22)
Emotional Invest. (Rel)	**EIR**	3.40	3.70	+0.31	**0.74** [0.46, 0.89]	**0.72**	**0.49**	**86.4%** (19/22)
Emotional Invest. (Mor)	**EIM**	3.84	3.82	-0.02	**0.56** [0.19, 0.80]	**0.58**	**0.39**	**100.0%** (22/22)
Social Causality	**SC**	3.39	3.59	+0.20	**0.70** [0.40, 0.87]	**0.63**	**0.41**	**100.0%** (22/22)
Aggression Management	**AGG**	3.68	4.14	+0.45	**0.53** [0.14, 0.78]	**0.44**	**0.59**	**90.9%** (20/22)
Self-Esteem	**SE**	3.86	3.43	-0.43	**0.73** [0.45, 0.88]	**0.59**	**0.45**	**95.5%** (21/22)
Identity & Coherence	**ICS**	4.26	4.11	-0.15	**0.51** [0.11, 0.76]	**0.49**	**0.42**	**95.5%** (21/22)
OVERALL COMPOSITE	**ALL**	**3.72**	**3.79**	**+0.07**	**0.66** [0.56, 0.73]	**0.65**	**0.45**	**94.9%** (167/176)

Notes on Table 2: Single-rater agreement reached pooled ICC(2,1) = 0.65, MAE = 0.45 points, and 94.9% within ≤ 1.0 point across the 176 paired ratings. Under Out-of-Fold LOOCV intercept calibration, agreement advanced to Pearson r = 0.70, ICC(2,1) = 0.70, and MAE = 0.47 points. Dimensional agreement varied systematically across scales. Affective and relational dimensions met or exceeded the training proficiency benchmark of ICC > 0.60 (AFF: ICC = 0.83; EIR: ICC = 0.72; SC: ICC = 0.63). In contrast, structural and impulse regulation scales showed more modest agreement (COM: ICC = 0.47; ICS: ICC = 0.49; AGG: ICC = 0.44; EIM: ICC = 0.58; SE: ICC = 0.59). Because the eight dimensions assess distinct psychological constructs, and because stories are clustered within protocols, the overall composite row serves as a descriptive development summary rather than an independent single-construct metric.

3.3 Psychometric Dispersion and Measurement Precision

The consensus scoring architecture preserves score variance without compressing ratings toward the scale midpoint. Across the 22 development cards, model score dispersion closely matched human expert variance (candidate SD = 0.76 vs. expert SD = 0.80, variance ratio = 0.90). Model ratings spanned the range from 1.5 to 7.0. A similar dispersion emerged in naturalistic clinical intake data (N = 39 cards, SD = 0.95). We calculated measurement error using Classical Test Theory formulas with human baseline variance (SD = 0.80). Under consensus scoring, the estimated Standard Error of Measurement is SEM = 0.47 points, yielding a Smallest Detectable Change of SDC_{95} = 1.31 points. Under intercept calibration, measurement error reduces to SEM = 0.44 points, with SDC_{95} = 1.21 points. These figures represent cross-sectional measurement models. Establishing certified thresholds

for longitudinal clinical change will require dedicated validation in repeated-measurement patient cohorts.

3.4 Known-Groups Construct Validity in Archival Cohorts

We examined construct differentiation in archival records (N = 73 subjects, k = 359 narratives) where independent SCORS-G ratings were not available. In the Character Pathology Benchmark (N = 50: n = 25 personality disorder cases vs. n = 25 matched controls), automated scores reflected expected clinical impairments. Clinical cases showed a moderate reduction in overall composite scores (M = 3.62 vs. 3.92, Cohen's d = 0.50). This reduction was concentrated in moral integration (d = 0.52), aggression regulation (d = 0.48), and self-esteem (d = 0.45). In the exploratory Schizophrenia Pilot Cohort (n = 3 subjects, 20 cards each, k = 60 cards vs. n = 62 controls, k = 179 cards), patients received lower ratings on causal coherence (M = 3.60 vs. 4.00, d = 1.35) and relational investment (M = 3.10 vs. 4.10, d = 0.86). However, because cards were repeated within a small group of three individuals, we adjusted for clustering at the subject level (df = 4.0). Under this adjustment, the group difference was not statistically significant (p = 0.099). These pilot observations indicate directional trends rather than validated diagnostic classification.

3.5 Latent Factor Structure Analysis (N = 1,074 Operational Cards)

We conducted a Principal Component Analysis with varimax rotation across 1,074 complete operational cards. The analysis extracted three components that together account for 79.2% of total variance (eigenvalues: 4.34, 1.10, 0.90). Component 1 accounted for 54.3% of the variance. It clustered cognitive and relational differentiation scales (COM, SC, EIR, and ICS, with loadings between -0.77 and -0.91). Component 2 accounted for 13.7% of the variance. It captured drive and superego regulation (AGG loading +0.92, EIM loading 0.59), while showing a near-zero loading on verbal complexity (-0.02). Component 3 accounted for 11.3% of the variance. It reflected affective valence and self-worth (SE loading -0.88, AFF loading -0.80). Component 1 forms a prominent general factor, and this extracted structure differs somewhat from human factor models reported in the literature (e.g., Stein et al., 2012; Siefert et al., 2018). Nevertheless, the results confirm that automated scores capture multidimensional variance, and that aggressive drive regulation operates independently from verbal complexity.

3.6 Stimulus "Card Pull" Profiles (N = 928 Cards across 11 Card Types)

We examined dimensional distributions across N = 928 operational cards spanning 11 stimulus types. Score patterns aligned with known stimulus pull characteristics described in the clinical and empirical literature (e.g., Ridenour et al., 2021; Siefert et al., 2016): Card 7GF elicited the lowest Affective Quality

(*M* = 3.42 vs. 4.21 grand mean, Δ = -0.79). Card 13MF elicited low Affective Quality (*M* = 3.81) alongside the lowest Aggression scores (*M* = 3.50), mirroring empirical findings where Card 13MF exerts sweeping negative pull across clinical and non-clinical cohorts (Ridenour et al., 2021). Card 8BM elicited pronounced aggressive themes (*M* = 3.62 vs. 3.97 grand mean, Δ = -0.35). Card 14 pulled for themes of solitary relational withdrawal (*M* = 2.39 vs. 4.11 grand mean, Δ = -1.72). In contrast, Card 17BM produced the highest overall composite (*M* = 4.83) and positive affect (*M* = 4.98), consistent with adaptive card pull dynamics observed on benevolent relational stimuli (e.g., Card 2 in Siefert et al., 2016). These systematic variations show that automated scores reflect expected stimulus pulls. Clinicians can use this baseline to separate standard card expectations from patient-specific projections.

3.7 Verbatim Evidence Grounding & Citation Audit

We audited quotation verification across 1,271 completed operational cards containing 5,912 citations. The substring verifier confirmed 90.75% of quotations (5,365 / 5,912) as continuous verbatim substrings of the source transcripts. When an excerpt failed verbatim verification, the system attached an advisory warning badge to that card. Across the dataset, 13.69% of cards (174 / 1,271) received this badge. To understand why unconstrained models misquote, we examined 1,672 unconstrained free-form citations. Before verification constraints were applied, only 28.23% matched verbatim. The remaining non-exact citations (*n* = 1,200) fell into two categories: linguistic variations (36.7%), which included spacing, punctuation differences, and Hebrew morphological prefixes; and generative paraphrases (63.3%), where the model rephrased the patient's words. Constrained decoding and automated substring checking highlight unverified excerpts in the dashboard, directing the clinician to inspect source text before incorporating findings into a report.

3.8 Test-Retest Scoring Stability Across Operational Runs

We evaluated test-retest reliability across 123 unique narratives scored in independent operational sessions. The repeated runs yielded an intraclass correlation of ICC(3,1) = 0.70, with an average Mean Absolute Difference (MAD) of 0.47 points. Across all paired ratings, 72.9% fell within ≤ 0.50 points of each other, and 44.1% produced identical scores. Aggression Management showed the highest stability, with a MAD of 0.34 points and 66.7% identical ratings across repeated runs. These results confirm consistent scoring reproducibility across independent automated sessions.

3.9 Multi-Card Battery Aggregation and the Battery Error Attenuation Law (*N* = 28 Cards, 224 Paired Ratings)

Single TAT cards elicit specific emotional reactions and stimulus pulls. In Generalizability Theory studies of the SCORS-G (Ridenour et al., 2021; Siefert et al., 2016), Person by Card interaction accounts for the largest proportion of variance in affective-relational dimensions (26% to 58%), whereas cognitive dimensions (COM, SC) are stable person-level traits (27% to 32%). Consequently, full assessment batteries aggregate signals across multiple cards, which mathematically dampens stimulus-specific interaction noise. To evaluate this principle, we scored four complete 7-card batteries (N = 28 cards, 224 paired ratings) from Stein & Slavin-Mulford (2018) against expert consensus. At the single-card level (N = 224 ratings), scores showed moderate correlation with expert consensus (Pearson r = 0.44, ICC = 0.40, MAE = 0.78 points), with 76.8% of ratings falling within ≤ 1.0 point. When scores were averaged to battery-level dimensional means (N = 32 composite means), agreement improved noticeably. Correlation rose to Pearson r = 0.62 and ICC = 0.50, while battery-level MAE decreased to 0.54 points. Across these composite means, 84.4% of dimensional composites and 100.0% of overall battery grand means fell within ≤ 1.0 point of expert consensus.

Table 3: Multi-Card Battery Benchmark and Longitudinal Trajectory Across 4 Clinical Protocols

← Scroll horizontally to view all columns →

Battery Protocol	Stimulus Cards	Consensus Grand Mean	Model Grand Mean	Grand Mean Delta (Δ)	Battery MAE	Pattern Correlation (r)
Ch. 12 Protocol 1	7 TAT cards	3.25	3.93	+0.68	0.71	0.35
Ch. 12 Protocol 2	7 TAT cards	3.62	3.68	**+0.05**	**0.36**	**0.84**
Ch. 15 Time 1 (Pre-Tx)	7 TAT cards	3.75	4.32	+0.57	0.59	**0.77**
Ch. 15 Time 2 (Post-Tx)	7 TAT cards	3.76	4.17	**+0.41**	**0.50**	**0.62**

Multi-card battery aggregation also supported clinical profile discrimination and longitudinal tracking. In Chapter 12, the system differentiated the severe traumatic borderline pathology of Protocol 1 (M = 3.25) from the organized neurotic constriction of Protocol 2 (M = 3.62 vs. Model M = 3.68, Δ = +0.05, pattern r = 0.84, Battery MAE = 0.36 points). In Chapter 15, the system tracked clinical changes across psychotherapy. Automated scores captured cognitive growth on Complexity and Social Causality (M = 2.86 → 3.21). At the same time, they reflected the patient's expanded emotional range on Affective Quality and Aggression Management, distinguishing healthy acknowledgment of past trauma from functional deterioration.

---

4. Discussion

4.1 Construct Validity and the Decision-Support Boundary

The empirical findings indicate that TAT Agent demonstrates meaningful psychometric sensitivity:

1. It captures known-groups differences in character pathology (d = 0.50) and exploratory pilot trends in reality testing (d = 1.35, p = 0.099).
2. It exhibits multi-dimensional variance across three components (79.2% variance explained), separating cognitive, impulse, and affective domains.
3. It reflects established stimulus card-pull variations across Murray TAT cards.

However, psychometric construct sensitivity must be distinguished from autonomous rater qualification. In formal SCORS-G training, single-rater proficiency requires ICC > 0.60 per dimension (Stein et al., 2011; Stein & Slavin-Mulford, 2018), while broader psychometric conventions benchmark good inter-rater agreement at ICC ≥ 0.70 (Cicchetti, 1994; Koo & Li, 2016). While TAT Agent's calibrated consensus architecture approaches this range (ICC(2,1) = 0.65–0.70), the system is designed strictly as an adjunctive decision-support aid operating under active clinician oversight.

4.2 The Baseline Comparison

A static midpoint baseline that predicts 4.0 for every dimension achieves an MAE of 0.65 points and 89.8% agreement within 1.0 point on training protocols. While TAT Agent improves upon this baseline (MAE = 0.45–0.47, ICC(2,1) = 0.65–0.70), this baseline illustrates that on restricted 7-point Likert scales with strong central tendencies, percentage agreement within 1.0 point is a generous metric. Furthermore, empirical human distributions rarely center at 4.0 across all dimensions (e.g., non-clinical grand mean ~3.8 in Bram, 2014; non-clinical undergraduate mean M = 3.51 in Siefert et al., 2016; outpatient composite M = 3.50, SD = 0.50 in Stein et al., 2018; DeFife et al., 2015). In a Hebrew-speaking community cohort, Handelzalts et al. (2014) observed adult dimension means spanning 3.23 to 4.02 with high rater agreement (ICC = 0.70–0.81), confirming that local Israeli baseline expectations align with international non-clinical distributions. Intraclass correlations, factor distributions, and known-groups effect sizes remain essential criteria for evaluating whether an automated system tracks genuine between-subject variance.

4.3 Evolution of Domain Boundary Constraints and Bias Mitigation

Developing the scoring constraints revealed several practical trade-offs:

1. **Unconstrained baseline:** Unconstrained models frequently hallucinated narrative details, compressed scores toward the scale midpoint (4.0), and exhibited affective halo bias. In particular, models assigned overly positive ratings to violent stories when those stories were written in fluent prose.
2. **Drive boundaries:** Enforcing strict scoring caps on destructive acts reduced scoring errors substantially. For example, capping overt homicide and unremorseful assault to Level 1–2 reduced Aggression MAE from 0.48 to 0.32 points. Similarly, bounding irrational survivor guilt to Level 3 improved moral investment scoring.
3. **Stimulus distance:** When patients lose psychological distance from the picture—such as addressing the examiner directly in anxiety—the model often over-interpreted this

disorganization as cognitive complexity. Requiring a Level 1 rating on Complexity for loss of distance corrected this tendency.

4. **Relational ceilings:** Models frequently over-interpreted dutiful compliance as mature mutual relationship. Introducing a ceiling of Level 2–3 for joyless obligation reduced this optimism bias, while targeted specialist review on high-conflict cards prevented the system from depressing scores on healthy family stories.

Together, these domain boundaries established practical guardrails against affective halo bias, while preserving scoring accuracy on adaptive narratives.

---

5. Clinical Decision-Support Safeguards

5.1 Decision-Support Workflow & Accountability

TAT Agent operates as a Clinical Decision-Support System (CDSS) structured around three core principles:

1. **Preliminary Drafting Aid:** The system generates provisional scores and citation references to reduce manual clerical scoring time.
2. **Citation Verification Telemetry:** Non-verbatim quotations are flagged directly in the dashboard, directing clinician attention to potential paraphrases.
3. **Mandatory Clinician Ownership:** The evaluating clinician retains full professional responsibility to review, adjust, or reject any rating. Diagnostic formulation and clinical synthesis remain the exclusive domain of the human examiner.

5.2 Automation Bias Countermeasures

To protect against automation bias, report generation modules include explicit notices:

> ⚠️ *CDSS Methodological Notice: Risk indicators and structural organization categories are heuristic screening aids. They do not constitute an autonomous diagnosis and require clinical verification via multi-method assessment.*

Clinicians are also reminded that a verified verbatim quotation confirms only that the words appear in the transcript; it does not guarantee that the model's score or clinical interpretation is valid.

---

6. Privacy & Deployment Architecture

1. **Local On-Premises Deployment:** The pipeline supports local open-weight models (e.g., DictaLM-24B, Llama-3, Mistral) hosted via local inference servers (LM Studio, vLLM). In this

configuration, narrative transcripts and scores are processed entirely on the local workstation without external network transmission.

2. **Cloud API Configuration:** When cloud inference endpoints are configured, communications utilize Zero Data Retention (ZDR) endpoints where available under provider agreements.
3. **Regulatory Design:** Technical workflows are architected with reference to Israel's Privacy Protection Law (5741-1981), the Patient's Rights Law (5756-1996), and GDPR principles for local data processing.

---

7. Methodological Limitations

1. **Development Sample Size (*N* = 22 Cards across 5 Protocols):** Prompt constraints were tuned using the 22 training narratives from Stein et al. (2011). These results reflect development calibration rather than broad out-of-sample rater equivalence.
2. **Small Clinical Subgroups:** The archival schizophrenia comparison was based on $n = 3$ subjects ($k = 60$ cards, with 20 repeated cards per subject). Due to small sample size and non-independence, subject-level differences did not reach statistical significance ($p = 0.099$), representing preliminary directional observations rather than clinical validation.
3. **Dimensional Variation on Specific Scales:** While affective and relational dimensions show single-rater agreement in the 0.65–0.83 range, cognitive, identity, and drive regulation scales (COM: ICC = 0.47; ICS: ICC = 0.49; AGG: ICC = 0.44; EIM: ICC = 0.58; SE: ICC = 0.59; SC: ICC = 0.63) exhibit lower agreement on the development set, underscoring the necessity of consensus scoring and active human clinical review.
4. **Exploratory Screening Heuristics:** The structural personality organization (NPO/BPO/PPO) heuristics described in Appendix B are unvalidated rule-based algorithms and should not be used for diagnostic categorization.
5. **Investigator Role:** The author is both a practicing clinical psychologist and the developer of the system. While calculations and scripts are documented for reproducibility, independent replication by third-party research teams is encouraged.
6. **Normative Generalizability Boundaries:** Although empirical findings in Hebrew-speaking adults (Handelzalts et al., 2014) demonstrate close alignment with international non-clinical norms (Bram, 2014), broad standardized normative distributions stratified across diverse age cohorts and clinical sub-populations in Israel remain an ongoing empirical requirement.

---

8. Declarations

8.1 Author Note

Ran Perlman, M.A., is an expert specialist clinical psychologist licensed by the Israel Ministry of Health (License No. 27-32174). Correspondence: `TAT_Online@ranperlman.com`.

8.2 Ethics

Secondary benchmarking of published training protocols (Stein et al., 2011) and de-identified archival clinical records does not involve direct human subject intervention. Archival cases were anonymized in compliance with institutional guidelines.

8.3 Data & Code Availability

The 22 clinical training narratives and reference ratings are documented in the *SCORS-G Training Manual* (4th ed., Stein et al., 2011; unpublished training manual), with theoretical rubrics published in Stein & Slavin-Mulford (2018). The canonical development benchmark results (`data/full_22card_consensus_benchmark.json`), live manual golden standard validation results (`data/scors_manual_validation_results.json`), replication bundle (`data/open_replication_bundle.json`), extracted normative baseline tables (`data/normative_sources/`), and reproduction scripts (`scripts/benchmark_dual_engine_reliability.py`, `scripts/run_manual_statistical_validation.py`, `scripts/run_open_replication_benchmark.py`) are maintained in the project repository.

8.4 Conflict of Interest

Ran Perlman is the developer of the TAT Agent Online platform and has a commercial interest in the platform. All empirical calculations and script outputs are archived for verification.

8.5 Funding

This work received no external grant or institutional financial support; it was independently developed and self-funded.

References

1. **American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014).** *Standards for educational and psychological testing*. American Educational Research Association.
2. **Bram, A. D. (2014).** Object relations, interpersonal functioning, and health in a nonclinical sample: Construct validation and norms for the TAT SCORS-G. *Psychoanalytic Psychology*, 31(3), 314–342. https://doi.org/10.1037/a0036662
3. **Cicchetti, D. V. (1994).** Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

4. **Dauphin, V. B., & Siefert, C. J. (2025).** From Llama to language: Prompt-engineering allows general-purpose artificial intelligence to rate narratives like expert psychologists. *Frontiers in Artificial Intelligence*, 8, 1398885. https://doi.org/10.3389/frai.2025.1398885
5. **DeFife, J. A., Goldberg, M., & Westen, D. (2015).** Dimensional assessment of self- and interpersonal functioning in adolescents: Implications for DSM-5's general definition of personality disorder. *Journal of Personality Disorders*, 29(2), 248–260. https://doi.org/10.1521/pedi_2013_27_074
6. **Dzega, A., Elyashar, A., Slobodin, O., Cohen, O., & Puzis, R. (2026).** Projective psychological assessment of large multimodal models using Thematic Apperception Tests. *arXiv preprint arXiv:2602.17108*.
7. **Handelzalts, J. E., Fisher, S., & Naot, R. (2014).** Object relations and real life relationships: A cross method assessment. *Scandinavian Journal of Psychology*, 55(2), 160–167. https://doi.org/10.1111/sjop.12102
8. **Hilsenroth, M. J., Stein, M. B., & Pinsker, J. (2007).** *Social Cognition and Object Relations Scale: Global Rating Method (SCORS-G)* (3rd ed.). Unpublished training manual.
9. **Kernberg, O. F. (1984).** *Severe personality disorders: Psychotherapeutic strategies*. Yale University Press.
10. **Koo, T. K., & Li, M. Y. (2016).** A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
11. **Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hrobjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011).** Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64(1), 96–106.
12. **Lingiardi, V., & McWilliams, N. (Eds.). (2017).** *Psychodynamic diagnostic manual: PDM-2* (2nd ed.). Guilford Press.
13. **Morgan, C. D., & Murray, H. A. (1935).** A method for investigating fantasies: The Thematic Apperception Test. *Archives of Neurology & Psychiatry*, 34(2), 289–306.
14. **Ridenour, J. M., Lewis, K. C., Siefert, C. J., Pitman, S. R., Knauss, D., & Stein, M. B. (2021).** Card pull effects of the Thematic Apperception Test using the Social Cognition and Object Relations-Global Rating Method on a complex psychiatric sample. *Clinical Psychology & Psychotherapy*, 28(2), 594–604. https://doi.org/10.1002/cpp.2554
15. **Siefert, C. J., Dauphin, V. B., et al. (2026).** From free association to free parameters: Machine-learning empowers artificial intelligence models to rate the affective-relational aspects of narratives like an expert psychologist. *Frontiers in Digital Health*, 8, 1762253. https://doi.org/10.3389/fdgth.2026.1762253
16. **Siefert, C. J., Stein, M. B., Slavin-Mulford, J., Sinclair, S. J., Haggerty, G., & Blais, M. A. (2016).** Estimating the effects of Thematic Apperception Test card content on SCORS-G ratings: Replication with a nonclinical sample. *Journal of Personality Assessment*, 98(4), 351–361. https://doi.org/10.1080/00223891.2016.1167696
17. **Siefert, C. J., Stein, M. B., Slavin-Mulford, J., Haggerty, G., Sinclair, S. J., Funke, D., & Blais, M. A. (2018).** Exploring the factor structure of the Social Cognition and Object Relations–Global

Rating Method: Support for two- and three-factor models. *Journal of Personality Assessment*, 100(2), 122–134. https://doi.org/10.1080/00223891.2017.1308785

18. **Stein, M. B., Hilsenroth, M. J., Pinsker-Aspen, J. H., & Primavera, L. H. (2009).** Validity of DSM-IV Axis V Global Assessment of Relational Functioning Scale (GARF): A multi-method assessment. *The Journal of Nervous and Mental Disease*, 197(1), 50–55. https://doi.org/10.1097/NMD.0b013e3181923e20
19. **Stein, M. B., Hilsenroth, M. J., Slavin-Mulford, J., & Pinsker, J. (2011).** *Social Cognition and Object Relations Scale: Global Rating Method (SCORS-G)* (4th ed.). Unpublished training manual.
20. **Stein, M. B., & Slavin-Mulford, J. (2018).** *Scoring stories and social cognition: The Social Cognition and Object Relations Scale-Global Rating Method (SCORS-G)*. Routledge.
21. **Stein, M. B., Slavin-Mulford, J., Sinclair, S. J., Siefert, C. J., & Blais, M. A. (2012).** Exploring the construct validity of the Social Cognition and Object Relations Scale in a clinical sample. *Journal of Personality Assessment*, 94(5), 533–540. https://doi.org/10.1080/00223891.2012.668594
22. **Stein, M. B., Slavin-Mulford, J., Sinclair, S. J., Siefert, C. J., Smith, M., Chung, W. J., Liebman, R., & Blais, M. A. (2015).** External validity of SCORS-G ratings of Thematic Apperception Test narratives in a sample of outpatients and inpatients. *Rorschachiana*, 36(1), 58–81. https://doi.org/10.1027/1192-5604/a000058
23. **Stein, M. B., Slavin-Mulford, J., Sinclair, S. J., Chung, W.-J., Roche, M., Denckla, C., & Blais, M. A. (2018).** Extending the use of the SCORS-G composite ratings in assessing level of personality organization. *Journal of Personality Assessment*, 100(2), 166–175. https://doi.org/10.1080/00223891.2016.1195394
24. **Stein, M. I. (1955).** *The Thematic Apperception Test: An introductory manual for its clinical use with adults* (Rev. ed.). Addison-Wesley.
25. **Westen, D. (1991).** Clinical assessment of object relations: Empirical research and the Social Cognition and Object Relations Scale. *Psychological Bulletin*, 109(3), 429–455.
26. **Westen, D. (1995).** *Social Cognition and Object Relations Scale: Q-sort for projective stories (SCORS-Q)*. Clinical manual. Cambridge, MA: Harvard University.

Appendix A: Partitioned Scoring Fallback Protocol

When complex clinical narratives trigger token deliberation limits under local reasoning models, the system activates an operational three-cluster fallback sequence:

1. **Cluster 1 (Cognitive-Structural Cluster):** COM, SC (USC), ICS.
2. **Cluster 2 (Relational-Affective Cluster):** AFF, EIR, SE.
3. **Cluster 3 (Impulse & Moral Cluster):** EIM, AGG.

Each cluster prompt receives full narrative access and operates under identical scoring constraints (§2.5).

Appendix B: Exploratory Decision-Support Heuristics (Unvalidated)

> **Clinical Disclaimer:** The structural personality organization (NPO/BPO/PPO) and thought-disorder screening heuristics described below are preliminary, rule-based exploratory algorithms. They **have not been validated** against external structured psychiatric criteria (e.g., PANSS, SCID-5, R-PAS). Furthermore, they rely on SCORS-G dimensions exhibiting modest single-rater reliability (Social Causality: ICC = 0.63; Identity Coherence: ICC = 0.49). They are presented strictly as heuristic aids and must not be used for autonomous diagnostic categorization.

B.1 Provisional Level of Personality Organization (Kernberg Framework)

Following Otto Kernberg's structural diagnosis framework (Kernberg, 1984) and PDM-2 (Lingiardi & McWilliams, 2017), protocol-level dimension means are synthesized into five structural axes, with quantitative thresholds centralized in the clinical constants registry:

1. **Reality Testing Axis:** Evaluated via Social Causality (*SC*). Intact reality testing requires *SC* ≥ 3.5, while scores below 2.5 indicate severe reality testing compromise.
2. **Identity Integration Axis:** Evaluated via the composite mean of Complexity of Representations (*COM*) and Identity Coherence (*ICS*). Consolidated identity requires a combined mean ≥ 4.5; moderate integration spans 3.7 ≤ mean < 4.5; scores between 2.8 and 3.69 denote structural vulnerability; and scores below 2.8 indicate identity diffusion.
3. **Defensive Organization Axis:** Evaluated via the balance of mature versus primitive defenses, requiring mature defenses ≥ 4.4 for neurotic organization, mixed defenses 3.6 ≤ mean < 4.4, and scores below 2.8 indicating pervasive reliance on primitive splitting.
4. **Superego & Moral Functioning Axis:** Evaluated via Moral Investment (*EIM*), with mature integration ≥ 4.4, conventional/moderate functioning 3.5 ≤ *EIM* < 4.4, and neurotic organization requiring *EIM* ≥ 4.0.
5. **Object Relations Quality Axis:** Evaluated via the composite mean of Emotional Investment (*EIR*) and Affective Quality (*AFF*), distinguishing mutual investment (≥ 4.4) and moderate investment (3.6 ≤ mean < 4.4) from need-gratifying (2.8 ≤ mean < 3.6) or alienated relations (< 2.8).

These axes determine provisional organizational levels:

- **Neurotic Personality Organization (NPO):** Concurrently intact reality testing (*SC* ≥ 3.5), integrated identity (combined mean ≥ 4.5), mature defenses (≥ 4.4), and integrated superego functioning (*EIM* ≥ 4.0).
- **Higher-Level Borderline Organization (Higher BPO):** Intact reality testing (*SC* ≥ 3.5), accompanied by moderate-to-vulnerable identity integration (2.8 ≤ mean < 4.5) and mixed defenses (≥ 2.8).

- **Lower-Level Borderline Organization (Lower BPO):** Borderline organization with identity diffusion (< 2.8), pervasive primitive defenses (< 2.8), impaired superego (< 2.8), or vulnerable reality testing (2.5 ≤ *SC* < 3.0).
- **Psychotic Spectrum Organization (PPO):** Severe reality testing collapse (*SC* < 2.5).

B.2 Structural Category Likelihood Indices

In addition to the five-axis structural tree, the platform computes two empirical category likelihood indices that capture distinct, orthogonal psychopathological dimensions:

1. **Borderline Organization Likelihood (0–100%):** Estimates the probability of borderline or externalizing character pathology across four weighted components: affective and moral volatility (standard deviation of *AGG* and *EIM*, 30%); critical floor drops denoting malignant aggression or primitive splitting (*AGG* or *EIM* ≤ 1.5, 30%); loss of stimulus distance (direct examiner appeals, 20%); and representational complexity deficit (*COM* mean < 3.35, calibrated against Bram, 2014 Table 2 non-clinical baseline, 20%). **Crucially, Borderline Organization Likelihood measures identity diffusion, primitive splitting, and affective volatility while fundamental reality testing remains intact.**
2. **Psychotic & Thought Disorder Likelihood (0–100%):** Estimates the probability of cognitive-perceptual or thought-disorder impairment across four weighted components: severe causal attribution breakdown (*SC* ≤ 2.5 alongside mean penalty floor, 35%); representational fragmentation (*COM* ≤ 2.5, 25%); narrative coherence collapse (*ICS* ≤ 3.5, 20%); and lexical pattern matching for bizarre or psychotic thought content (somatic fragmentation, remote control, or ungrounded persecutory ideation, 20%). **Crucially, this index measures the specific breakdown of reality testing, boundaries, and logical coherence.**

Clinical Orthogonality & Benchmark Separation

Because Borderline Personality Organization (BPO) and Psychotic Organization (PPO) represent structurally distinct diagnostic categories in psychodynamic theory (Kernberg, 1984), these two likelihood meters diverge predictably across clinical populations:

- **Normative Non-Clinical Controls:** Borderline Likelihood 0%–5%, Psychotic Likelihood 0%–3% (intact identity integration and intact reality testing).
- **Severe Personality Disorders (BPO):** Borderline Likelihood 50%–83% (elevated character pathology, splitting, affective volatility), while Psychotic Likelihood remains 0%–15% (**reality testing remains intact**).
- **Psychotic & Organic Spectrum (PPO):** Psychotic Likelihood 35%–72% (**severe breakdown in reality testing**, cognitive slippage, or organic cognitive impairment), with secondary character volatility.

An elevated Borderline Organization Likelihood does **not** indicate psychosis or loss of reality testing. Any finding of elevated likelihood requires systematic multi-method assessment with established

performance-based instruments (e.g., R-PAS Thought and Perception Index) and comprehensive clinical interview.