Cheryl Burris
EDIT 891-2
7 May 2026

Evaluation Protocol Infrastructure

STATUS: Not a full evaluation design. Paired with: Failure Taxonomy v2 (4 May 2026) and DDR-21 (7 May 2026).

1. What still needs to be evaluated

A test to evaluate a single central hypothesis: The scaffolds, bounded and designed correctly, will support the regulatory processes they appear to replace rather than substituting for them permanently.

This hypothesis generates three operational questions:

Q1: Does the planner produce behavioral change in planning and regulation, not just usage? This distinguishes session engagement (L1 territory) from functional change (L2 territory).

Q2: If behavioral change occurs, is it in the direction of increased self-regulatory capacity or increased scaffold dependency? This is the L3 failure boundary.

Q3: Do the ADHD and ASD paths produce these effects through their respective mechanisms, or are the mechanisms confounded? ADHD arousal-driven variability (Isaac et al., 2024) and ASD context-driven variability (Mantzalas et al., 2022; Lage et al., 2024) must be evaluated with population-specific criteria.

2. Failure Mode Operationalization

Source: Failure Taxonomy v2 (4 May 2026). All indicators are provisional pending the evaluation design. Priority is given to behavioral indicators that can be captured from the planner's existing data output (session logs, CSV export) without requiring additional self-report instruments.

Level	Definition	ADHD-path indicator	ASD-path indicator
L1	Planner used within a session but produces no usable plan.	Task entry initiated; partial entries only	Session completed and plan produced,

Session failure	Behavioral indicators (either path): • Zero tasks scheduled after completed On-Boarding. • AI Weekly Prep output generated; no tasks entered into This Week. • Orphan prompt dismissed across three or more consecutive sessions. • Week load score = 0 pts with empty task list for the viewed week.	(energy selected, no due dates). Session abandoned mid-entry.	but plan is rigid with no buffer. It cannot survive first unplanned interruption. The plan exists, but the recovery path does not.
L2 Functional failure	Planner used correctly across multiple sessions; no measurable change in planning behavior. Behavioral indicators (either path): • Estimate-actual time gap shows no convergence after four or more weeks. • Over-commitment persists: hours scheduled exceeds available hours across two or more consecutive weeks. • System abandoned after two weeks with no Off-Boarding logged. • Self-efficacy pulse holds flat or trends downward across four or more sessions.	Forethought phase completed reliably. The planner is not consulted during Performance. The plan exists, but the execution loop does not close.	Planner followed exactly, but unpredictable course environment (late assignments, rubric changes) violates the plan session by session. Distress escalates. Not a user failure, but a design scope problem.
L3 Design failure	Planner functions correctly; behavioral change is in the wrong direction. Behavioral indicators (either path): • AI suggestion acceptance rate exceeds 80% without revision across three or more sessions (Weekly Prep Ritual, task chunking, Agenda suggestions).	Urgency scaffolds (Pomodoro, priority prompts) produce compliance during sessions. Planning collapses when planner is unavailable. Scaffold substituted for regulation.	Planner structure becomes a safety object. Any deviation from planner output triggers distress. Calm Mode required as a precondition rather than used selectively.

	• Self-efficacy pulse declining despite regular use and apparent task completion. • Absence of planner triggers dysregulation. Inferred from session gaps followed by crisis-entry patterns (Recovery flow immediately after multi-day gap). • Off-Boarding reflection specificity declining over time, with shorter, less detailed entries as a proxy for agency erosion.		

3. Measurement Validity Threats

These threats were identified in Failure Taxonomy v2 and DDR-21. Each must be addressed in the evaluation design before data collection begins. An unmitigated threat does not block evaluation. It must be named as a methodological limitation if it cannot be fully mitigated.

#	Threat	Description	Mitigation
T1	**calmMode preference override** (ADHD path, L3 risk)	Users may habitually select Scattered at On-Boarding not because they are scattered but because they prefer the calm interface. Once this pattern is established, calmMode auto-enables every session and cannot be disabled mid-session. The check-in stops measuring arousal state and starts functioning as a preference override. Self-reported check-in state becomes unreliable as a data point.	Cross-check Scattered self-report against behavioral indicators: task entry rate, session length, and Pomodoro completion rate. If Scattered is reported consistently but behavioral indicators are congruent with Focused sessions, flag as a preference pattern rather than a state report. Do not treat check-in state as a reliable independent variable without this triangulation.
T2	**AI suggestion acceptance as agency proxy** (Both paths, L3 risk)	High AI suggestion acceptance rate could mean three distinct things: (a) the user genuinely agrees with the suggestion; (b) the user lacks metacognitive capacity to evaluate it; or (c) the user accepts it to reduce decision demand.	Pair acceptance rate with Off-Boarding reflection quality. A user who accepts AI suggestions but produces specific, self-authored reflection language is not exhibiting agency erosion. A user whose reflections become shorter and more generic over time

		Only interpretation (b) and (c) represent failure. The acceptance rate alone cannot distinguish them.	while acceptance rate rises is a candidate for L3 monitoring. ASD-specific: also check whether Agenda-view AI acceptance correlates with Heavy sensory load check-in. Predictability-seeking in high-demand states is distinct from habitual deference.
T3	**Self-report reliability** (Both paths)	Self-efficacy pulse and Off-Boarding reflection are subject to social desirability effects. Users may report improving self-efficacy because they feel they should, not because their regulatory capacity has changed.	Triangulate self-report with behavioral usage data. If self-efficacy scores rise while task completion rate, estimate-actual calibration, and Pomodoro follow-through remain flat, flag the divergence as a methodological note rather than treating the self-report as valid.

4. Population-Specific Success Criteria

The ADHD and ASD paths were designed around structurally distinct mechanisms. They cannot be evaluated with identical success criteria. The following criteria are provisional and designed to be tested, not assumed.

ADHD path (arousal-driven variability)	ASD path (context-driven variability)
Success indicators: • Estimate-actual calibration gap narrows over four or more weeks. • Task completion rate higher on days when Forethought phase was completed vs. days when it was skipped. • Orphan task accumulation rate decreases over time. • Energy level log shows diversity across sessions and not always High. Accurate self-assessment, not optimistic reporting. • Self-efficacy pulse improves alongside behavioral task completion (not divergent). L3 watch indicators:	Success indicators: • Session completion rate stable or improving; fewer sessions ending in abandonment. • Recovery and Re-Entry flow frequency decreasing over time; fewer sessions requiring re-entry after overload. • Calm Mode use is selective: not applied at every session regardless of check-in state. Evidence of state-responsive use. • Pace/Rescope modal used without session abandonment; distress tolerance in response to plan deviation improving. • Self-efficacy pulse stable or improving; does not diverge from behavioral indicators. L3 watch indicators:

• Urgency scaffolds used in every session without exception; no sessions where user initiates without Pomodoro or priority prompts. • Planning collapses during sessions where On-Boarding was skipped since this is a planner-dependent initiation.	• Calm Mode required at every session regardless of sensory load check-in result. • AI Agenda suggestions accepted at >80% without revision across three or more sessions. • Session gaps of two or more days followed immediately by Recovery flow entry.

5. Data Sources

The planner's CSV export (Progress page) provides the primary behavioral data stream. No additional instrumentation is required for the behavioral indicators listed above. Self-report data is captured by existing Off-Boarding features.

Data source	Variables captured	Relevant to
Session logs (On-Boarding completion)	• Check-in state (energy, adhdCogLoad, sensoryLoad) • Effective interface mode • Session timestamp	T1 mitigation (calmMode preference check) L1 indicators Population path identification
Task data (CSV export)	• Estimate vs. actual time • Completion status • Due date vs. completion date • Orphan status (undated tasks) • Priority level • depletedFlag	L1, L2 indicators Estimate-actual calibration Over-commitment detection Orphan accumulation rate
AI feature usage logs	• Weekly Prep output generated (Y/N) • Tasks entered after Prep (count) • Task chunking invocations • Agenda AI suggestions presented / accepted / revised	T2 mitigation (acceptance rate proxy) L3 indicators - agency erosion detection
Pomodoro usage	• Timer completion rate per session • Minimize/expand frequency (V6-2)	L2 ADHD-v (execution loop closure) L3 scaffold dependency BS-5 fading signal
Off-Boarding logs	• Self-efficacy pulse score • Mastery experience rating	T3 mitigation (triangulation with behavioral data)

	• Phase label (Green/Yellow/Red) • Reflection notes (text)	L2 and L3 self-report indicators Reflection specificity proxy for agency
Recovery and Re-Entry flow events	• Re-entry screen completions • Session gap preceding re-entry (days)	ASD-path success indicator L3 ASD-v watch indicator

6. What the Next Phase Cannot Claim

This section serves as the methodological infrastructure for the next phase of development. The next phase is a DBR evaluation cycle, not a randomized controlled trial. The following claims will be out of scope regardless of what the data shows:

- **Causal efficacy.** The next phase cannot establish that the planner caused observed behavioral changes. Confounds are uncontrolled. The claim available is: The design is consistent with the hypothesized mechanism, or it is not.
- **Generalizability.** Findings apply to the specific population, context, and version evaluated. Cross-population generalization requires separate evaluation cycles.
- **Validated measurement.** The check-in instruments (adhdCogLoad, sensoryLoad) are design constructs, not validated psychometric measures. Their outputs are design inputs, not clinical assessments.
- **Stable user population.** ADHD and ASD users' states vary within and across sessions. Behavioral data reflects a moving target. Point-in-time comparisons are weaker than longitudinal trend analysis.

The claim the next phase can support: The planner's design decisions are responsive to the evidence base; the failure modes identified in the last phase were or were not observed in use; and the design hypotheses generated in the last phase require revision, extension, or further testing in subsequent cycles. That is the DBR contribution.

References

Isaac, V., Lopez, V., & Escobar, M. J. (2024). Arousal dysregulation and executive dysfunction in attention deficit hyperactivity disorder (ADHD). *Frontiers in Psychiatry, 14*, Article 1336040. https://doi.org/10.3389/fpsyt.2023.1336040

Lage, C., Smith, E. S., & Lawson, R. P. (2024). A meta-analysis of cognitive flexibility in autism spectrum disorder. *Neuroscience and Biobehavioral Reviews*, *157*, Article 105511. https://doi.org/10.1016/j.neubiorev.2023.105511

Mantzalas, J., Richdale, A. L., Adikari, A., Lowe, J., & Dissanayake, C. (2022). What Is autistic burnout? A thematic analysis of posts on two online platforms. *Autism in Adulthood*, *4*(1), 52–65. https://doi.org/10.1089/aut.2021.0021