FOYL RESEARCH · TECHNICAL REPORT TR-2026-01

An Instrumented Simulation Testbed for Measuring Cognitive Load and Decision-Making in Security Operations

Aidan Hassell · foyl Research (independent) · research.foyl.io
Correspondence: support@foyl.io · August 2026 · Working paper, v1.0

ABSTRACT

Security operations center (SOC) analysts make rapid, consequential decisions in high-noise, high-tempo environments, yet most security training and tooling is evaluated on whether it functions rather than on how it shapes human performance. This report describes **foyl Research**, a web-based simulation testbed that treats a working analyst environment as a research instrument. The platform renders realistic triage and phishing-detection tasks as controlled experiments: it manipulates signal-to-noise ratio, time pressure, and motivational framing between participants while logging a millisecond-grained interaction stream. From that stream it derives signal-detection measures (sensitivity d' and decision criterion c), time-to-first-diagnostic, and friction (hesitation, backtracking, context-switching), alongside validated workload (NASA-TLX) and security self-efficacy instruments. The system is anonymous by construction, stores no personally identifying information, and produces analysis-ready, exportable data. We present the architecture, the instrumentation, the factorial experimental design, and a pre-registered evaluation protocol. The contribution is a reusable, low-friction workbench for empirically studying how interface and environmental factors move analyst attention, workload, sensitivity, and speed.

1. Introduction and motivation

SOC analysts triage a stream of alerts that is dominated by false positives; qualitative studies report false-positive rates so high that genuine threats are routinely missed or dismissed, and analysts describe chronic alert fatigue and burnout [8, 10]. The gap between sanitized training material and the ambiguous, noisy conditions of production work is itself a human-factors problem: instruction that never exposes learners to realistic signal-to-noise ratios does not build the decision calibration the job demands [5, 9]. Yet the field lacks accessible instruments for measuring how analysts actually decide under these conditions.

This work asks a measurement question rather than a tooling one: *how do environmental and interface factors change an analyst's sensitivity, decision bias, workload, and speed?* To study it empirically we built an instrument that (i) reproduces ecologically grounded analyst tasks, (ii) manipulates the conditions of interest under experimental control, and (iii) captures behavior at a resolution that supports signal-detection and workload analysis. The design is deliberately

deployable to non-expert participants at scale, anonymous, and free of institutional-review friction for low-risk pilot work.

2. Related work

Signal detection in security decisions. Triage and phishing detection are naturally framed as detection tasks, in which an analyst separates a signal (a genuine threat) from noise (benign activity). Signal-detection theory (SDT) decomposes performance into sensitivity, d', and a criterion, c, that are independent of one another [1, 2]; this separation is essential because two analysts with identical accuracy may differ sharply in how readily they escalate. We adopt the log-linear correction of Hautus [12] for extreme hit and false-alarm proportions.

Workload and cognitive load. The NASA Task Load Index [3] provides a validated multidimensional measure of perceived workload; cognitive-load theory motivates the expectation that noise and time pressure raise extraneous load and degrade performance [11].
Human-in-the-loop security. Cranor's framework [5] situates the analyst as a component whose failures are predictable from communication, capability, and motivation; folk-model and usability work shows that non-expert mental models drive security behavior [6, 7]. **Self-efficacy.** Bandura's construct [4] predicts persistence and performance under difficulty and motivates our framing manipulation and pre/post self-efficacy measure.

3. System architecture

The instrument is a three-stage signal chain (Figure 1 shows the detection model it measures). A **simulation engine** renders analyst tasks in a self-consistent fictional enterprise; a **telemetry pipeline** captures every participant action; and a **condition engine** parameterises the task each participant receives.

- **Simulation engine.** Two task families are implemented: an alert-triage queue (escalate / dismiss / tune) and a phishing-inspection task (flag red-flag elements, then quarantine / release). All entities, incidents, and infrastructure are fictional and internally consistent.
- **Telemetry pipeline.** An append-only, idempotent event stream records time-stamped actions keyed to an anonymous, client-minted participant code. Events buffer locally and flush in batches, tolerating intermittent connectivity without loss or duplication.
- **Condition engine.** A pure function of a seed and the assigned factor levels produces the specific stimulus set, so any condition cell is reproducible and URL-addressable for directed runs.

The platform is a static front end with serverless functions over a lightweight relational store. It records no email, name, IP address, or user agent; the only identity is an anonymous code held in the participant's browser, and a short-lived salted hash of network metadata is used solely for abuse throttling and discarded within a day. Aggregate results and raw tables are exportable as CSV/JSON for offline analysis in R, Python, or SPSS.

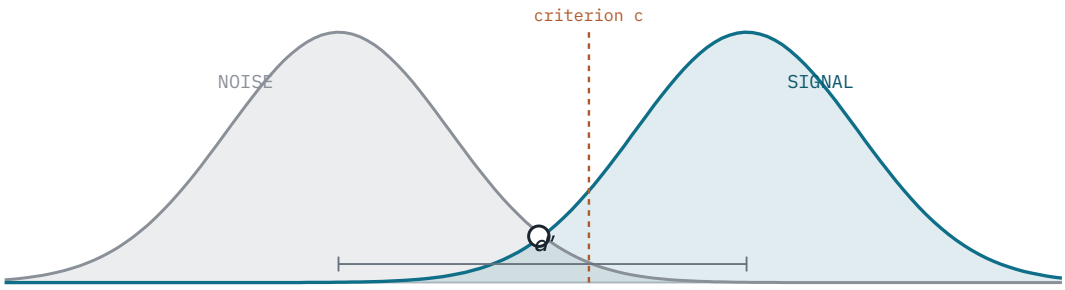

Figure 1. The signal-detection model the instrument measures. Sensitivity (*d′*) is the standardized separation of the signal and noise distributions; the criterion (*c*) is the analyst's escalation threshold. Both are recovered per participant per condition from hits and false alarms.

4. Instrumentation and measures

Each task yields behavioral measures computed from the event stream, plus self-report instruments administered at fixed points. Measures are stored per participant per task per condition.

SYMBOL	MEASURE	DEFINITION
d'	Sensitivity	z(hit rate) − z(false-alarm rate), with Hautus log-linear correction [12].
c	Criterion	−½[z(hit rate) + z(false-alarm rate)]; escalation bias.
t_0	Time-to-first-diagnostic	Latency from stimulus onset to the first committed decision.
RT	Response time	Median per-decision latency across the task.
∫	Friction	Hesitation (dwell over threshold), verdict backtracking, and context/window switches.
TLX	Workload	Six NASA-TLX subscales rated immediately after each task [3].
SE	Self-efficacy	Five-item security self-efficacy scale, administered pre and post [4].

5. Experimental design

The core design is a 2×2×2 between-subjects factorial over three manipulated environmental variables, with participants randomly assigned to a condition cell on session start; every cell is also URL-addressable so a researcher can direct assignment for balanced or targeted runs.

FACTOR	LEVELS	MANIPULATION	PRIMARY HYPOTHESIS
Signal noise	`low / high`	Decoy entities, distractor log lines, and queue length raise the signal-to-noise ratio.	High noise lowers *d′* and raises workload.
Time pressure	`off / on`	A per-decision countdown imposes a deadline.	Pressure shifts the speed–accuracy tradeoff and criterion.
Framing	`control / mastery`	Feedback wording emphasises penalties versus growth/agency.	Mastery framing raises post-task self-efficacy.

Dependent measures are the behavioral and self-report variables of Section 4. Because sensitivity and criterion are separated, the design can distinguish a genuine change in discrimination from a mere shift in escalation bias, which raw accuracy conflates.

6. Planned evaluation

An initial pilot (target N ≈ 40, convenience sample of students and practitioners) will establish task reliability, instrument sensitivity, and effect-size estimates for a subsequent powered study. The analysis plan treats condition as the between-subjects factor and analyses *d′*, *c*, time-to-first-diagnostic, and NASA-TLX by factorial ANOVA, with self-efficacy analysed as a pre/post change by framing. Data collection is anonymous and low-risk; the platform's export produces tidy, participant-level tables suitable for mixed-effects modeling as sample size grows. *This report describes the instrument and protocol; no participant results are reported here.*

7. Limitations and future work

The tasks are simulations, and generalization to live SOC work is an empirical question the pilot begins to address; convenience sampling limits external validity; and truth values are, by design, teaching-visible in some surfaces and must be handled carefully as an experimental control. Planned extensions include a risk-perception / mental-models elicitation, a security-nudge A/B module, and a human–AI reliance task in which a sometimes-incorrect assistant recommendation measures over- and under-reliance. The instrument is built to accept these as additional task modules without changes to the data pipeline.

References

[1] Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* Wiley.

[2] Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide* (2nd ed.). Lawrence Erlbaum.

[3] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX: Results of empirical and theoretical research. *Advances in Psychology,* 52, 139–183.

[4] Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review,* 84(2), 191–215.

[5] Cranor, L. F. (2008). A framework for reasoning about the human in the loop. *Proc. UPSEC.*

[6] Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Why phishing works. *Proc. ACM CHI.*

[7] Wash, R. (2010). Folk models of home computer security. *Proc. SOUPS.*

[8] Sundaramurthy, S. C., et al. (2016). Turning contradictions into innovations, or: How we learned to stop whining and improve security operations. *Proc. SOUPS.*

[9] Sasse, M. A., Brostoff, S., & Weirich, D. (2001). Transforming the 'weakest link' — a human/computer interaction approach to usable and effective security. *BT Technology Journal,* 19(3), 122–131.

[10] Alahmadi, B. A., Axon, L., & Martinovic, I. (2022). 99% false positives: A qualitative study of SOC analysts' perspectives on security alarms. *Proc. USENIX Security.*

[11] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science,* 12(2), 257–285.

[12] Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of *d′*. *Behavior Research Methods,* 27(1), 46–51.