Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 2491–2497
https://doi.org/10.54941/ahfe1008474

Detecting Silent Automation Reliability Degradation—A Signal Detection Theory Analysis

Cheng Tan and Liang Ma

Department of Industrial Engineering, Tsinghua University, Beijing, China

ABSTRACT

Silent reliability degradation provides no explicit alert, requiring operators to distinguish a genuine system-state change from ordinary stochastic errors. Using the previously reported first-report dataset (N = 60; 720 participant-scenario trajectories), this secondary signal detection theory (SDT) analysis formally decomposed terminal failure-report behavior into discrimination sensitivity (d-prime) and relative decision criterion (c). Reliability losses of 5% and 10% were treated as signal-present states, 0% loss as signal absent, and any first failure report within 100 formal monitoring trials as a positive response. Larger reliability loss was associated with higher d-prime. Higher initial reliability was associated with both higher d-prime and a more conservative c. Thus, the slower yet more accurate pattern previously observed under high initial reliability is consistent with the joint contribution of better terminal state discrimination and more conservative reporting. This analysis quantitatively complements prior time/accuracy and psychometric-threshold analyses of the same dataset.

Keywords: Automated decision aids, Silent reliability degradation, Signal detection theory, Sensitivity, Decision criterion

INTRODUCTION

Automated decision aids are increasingly used in safety-critical inspection, transportation, and clinical-support tasks. Their nominal accuracy is important, but safe performance also depends on whether operators can recognize when actual capability no longer matches the reliability established earlier in interaction. When a system degrades without an explicit warning, isolated errors are ambiguous: they may reflect ordinary stochastic variability or a persistent change in the underlying reliability state. This ambiguity makes silent reliability degradation a supervisory detection problem, not merely a count of automation errors (Lee and See, 2004; Parasuraman and Riley, 1997).

Signal detection theory (SDT) provides a principled way to separate two components of such judgments. Sensitivity, d-prime, summarizes how well responses discriminate signal-present from signal-absent states. The relative criterion, c, summarizes the tendency to issue a positive response after accounting for the locations of the two response distributions. Higher d-prime indicates

Received September 3, 2026; Revised October 9, 2026; Accepted October 14, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

better state discrimination, whereas positive c indicates a more conservative failure-report policy and negative c a more liberal policy (Macmillan and Creelman, 2005; Stanislaw and Todorov, 1999). This distinction matters because fewer false alarms can reflect better discrimination, a stricter reporting policy, or both.

The source experiment examined first reports of silent reliability loss in a smartphone-screen scratch-inspection task. Larger reliability losses produced earlier and more accurate failure judgments, whereas 90% rather than 75% initial reliability produced a slower but more accurate pattern (Tan et al., 2026a). That study also invoked SDT qualitatively to interpret perception accuracy, proposing that a more stable high-reliability baseline could improve state discrimination while a conservative reporting tendency could reduce false alarms. A subsequent psychometric secondary analysis characterized loss-to-report functions across monitoring windows and derived a loss-based operational detection threshold at the 50% report criterion (Tan et al., 2026b). However, neither analysis estimated d-prime and c from the terminal report/no-report response structure.

Accordingly, the present analysis does not seek to re-establish the previously reported condition effects. Instead, at the 100-trial endpoint it formally quantifies the proposed SDT account. Within each initial-reliability condition, 0% loss defines the noise state and 5% and 10% loss define two signal strengths. We ask how terminal d-prime varies across loss and initial reliability, whether higher initial reliability is associated with a more conservative relative c, and how the two quantities jointly characterize the previously observed speed-accuracy pattern.

METHOD

Data Source, Provenance, and Design

The present study is a secondary analysis of the dataset originally reported by Tan et al. (2026a); no new participant data were collected. The same dataset was later analyzed psychometrically by Tan et al. (2026b). Here the estimand is different: terminal report/no-report behavior is decomposed into SDT sensitivity and criterion. The original within-subject experiment included 60 participants and manipulated reliability loss (0%, 5%, and 10%), initial reliability (75% and 90%), and automation error type across 12 scenarios. Each scenario contained a training phase and a 100-trial formal monitoring phase. Participants verified individual automation outputs while monitoring the system as a whole and pressed a Failure button when they first judged that sustained reliability degradation had occurred.

For this endpoint analysis, a trajectory was coded positive if a first failure report occurred within the 100 formal trials and negative otherwise. Because the present SDT question concerns system-level degradation rather than error-type differences, the two automation error-type conditions (false-alarm errors and misses) were pooled within each initial-reliability by loss cell,

yielding 120 trajectories per cell. This pooling is analytical and does not imply psychological equivalence of the two error types.

Signal-Detection Measures

Signal present was defined as a genuine reliability decrease (5% or 10% loss), and signal absent as no decrease from the trained baseline (0% loss). A failure report in a signal-present trajectory was a hit; no report was a miss. A report in a signal-absent trajectory was a false alarm; no report was a correct rejection. Rates were computed separately for the 75% and 90% initial-reliability conditions, so each loss level was compared with the no-loss condition sharing the same initial reliability.

Because each participant contributed only two trajectories per initial-reliability by loss cell, individual d-prime and c estimates would be unstable. Group-level descriptive SDT summaries were therefore computed as d-prime = $z(H) - z(FA)$ and $c = -0.5[z(H) + z(FA)]$, where H and FA are hit and false-alarm rates. No rate was 0 or 1, so no continuity correction was required. Reported contrasts are descriptive decompositions, not inferential significance tests.

RESULTS

Across the 720 trajectories, participants reported failure in 413 and made no report in 307. Under 75% initial reliability, the no-loss false-alarm rate was 0.375; hit rates were 0.700 at 5% loss and 0.792 at 10% loss. Under 90% initial reliability, the false-alarm rate was 0.117, while hit rates were 0.642 and 0.817. Table 1 provides the response counts used in the SDT decomposition.

Table 1: Pooled endpoint report counts by initial reliability and reliability loss.

Initial Reliability	Loss	SDT Role	Reports	Trials	Report Rate
75%	0%	Noise	45	120	0.375
75%	5%	Signal	84	120	0.700
75%	10%	Signal	95	120	0.792
90%	0%	Noise	14	120	0.117
90%	5%	Signal	77	120	0.642
90%	10%	Signal	98	120	0.817

Note. Counts are re-tabulated from the dataset reported by Tan et al. (2026a) solely to define the present terminal SDT response structure; they are not new observations. The two automation error-type conditions were pooled.

Discrimination Sensitivity

At both initial-reliability levels, d-prime was higher at 10% than at 5% loss (Figure 1; Table 2). Under 75% initial reliability, d-prime was 0.843 at 5% loss and 1.131 at 10% loss. Under 90% initial reliability, the corresponding values were 1.555 and 2.095.

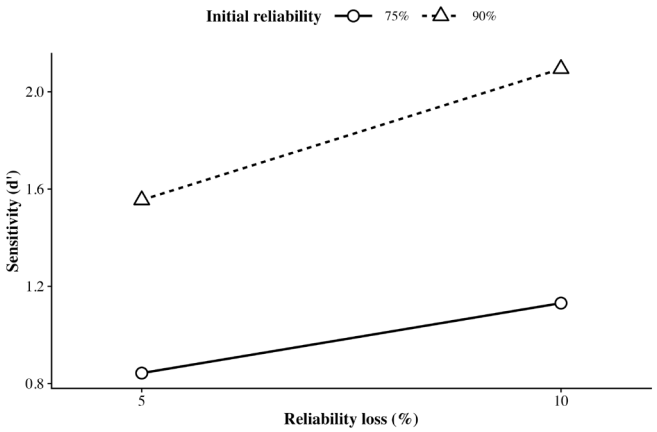


Figure 1: Sensitivity (d-prime) by reliability loss and initial reliability.

Table 2: Sensitivity estimates and descriptive contrasts.

Section	Condition / Contrast	Estimate
Condition	75% / 5%	0.843
Condition	75% / 10%	1.131
Condition	90% / 5%	1.555
Condition	90% / 10%	2.095
Contrast	Loss: 10% - 5%	0.414
Contrast	Initial: 90% - 75%	0.838
Contrast	Interaction	0.252

Relative Decision Criterion

Criterion estimates are shown in Figure 2 and Table 3. At 75% initial reliability, c was -0.103 at 5% loss and -0.247 at 10% loss. At 90% initial reliability, the corresponding values were 0.414 and 0.145. Thus, the terminal response pattern was more conservative under high initial reliability at both signal strengths.

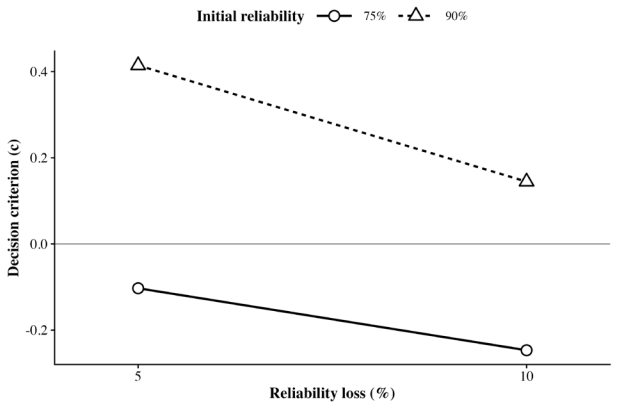


Figure 2: Relative decision criterion (c) by reliability loss and initial reliability. Positive values indicate a more conservative failure-report policy.

Table 3: Relative decision criterion estimates and descriptive contrasts.

Section	Condition / Contrast	Estimate
Condition	75% / 5%	–0.103
Condition	75% / 10%	–0.247
Condition	90% / 5%	0.414
Condition	90% / 10%	0.145
Contrast	Loss: 10% - 5%	–0.207
Contrast	Initial: 90% - 75%	0.454
Contrast	Interaction	–0.126

Note. Loss and interaction contrasts in c are algebraically dependent on the corresponding d-prime contrasts because signal conditions share one false-alarm baseline.

Joint Pattern

Across the two SDT summaries, 90% rather than 75% initial reliability corresponded to a descriptive d-prime contrast of 0.838 and a conservative shift in c of 0.454. The 10%-5% loss contrast was 0.414 for d-prime and –0.207 for c; the corresponding difference-of-differences were 0.252 and –0.126. These values describe the endpoint response structure and should not be interpreted as independent significance tests.

DISCUSSION

This secondary analysis had a deliberately narrow purpose: to formalize an SDT interpretation of a previously reported behavioral pattern, not to re-test whether reliability loss or initial reliability affected perception time or accuracy. At the 100-trial endpoint, larger loss corresponded to higher d-prime, whereas high initial reliability combined higher d-prime with a more positive, conservative c. The previously observed slower-but-more-accurate pattern under high initial reliability can therefore be described at the endpoint as a joint pattern of better state discrimination and greater reluctance to declare failure.

Higher d-prime under 90% initial reliability is consistent with a baseline-quality account: more reliable training experience may support a more stable representation of expected system performance, making no-loss and degraded states easier to distinguish. This interpretation also complements the psychometric finding that high initial reliability altered the loss-to-report function and maintained a higher operational detection threshold in later monitoring windows (Tan et al., 2026b). Importantly, the present estimates remain behavioral summaries and do not identify latent sensory or neural distributions.

Criterion results require disciplined interpretation. Conventional c is defined relative to the midpoint of the signal and noise distributions. When two signal strengths share one noise distribution, changing signal strength necessarily changes both d-prime and c even if an absolute response boundary is unchanged. Accordingly, the negative loss contrast in c is not independent evidence that larger degradation lowered an internal threshold.

The initial-reliability contrast is more informative because both hit and false-alarm rates differ, but even here c is a relative response-policy index. It should also not be equated with the loss-based operational detection threshold reported by Tan et al. (2026b): that threshold is expressed in percentage points of reliability loss at a specified monitoring window, whereas c is a dimensionless SDT location measure of terminal responses.

The results imply two complementary design targets. Interfaces can improve discrimination by making performance trends, recent error history, and deviations from the trained baseline more observable. They can support criterion calibration by communicating the costs of delayed reports and false alarms and by providing graded reliability states or confidence-qualified escalation rather than a single binary failure button. A system designed only to reduce false alarms may inadvertently promote delayed intervention; a system designed only to increase alerts may promote indiscriminate responding. Human-centered monitoring support should improve evidence quality while aligning the reporting policy with operational risk (Lee and See, 2004).

Three limitations define the scope of the conclusions. First, this is a secondary analysis of the same source dataset and therefore an analytical extension, not an independent replication. Second, because per-participant cell counts were sparse, the estimates are group-level descriptive SDT summaries and cannot replace individual-level or hierarchical latent SDT models. Third, the analysis uses only the 100-trial endpoint and cannot describe within-scenario evolution of sensitivity or criterion. The conclusions are therefore limited to the terminal response structure represented by the present operationalization.

CONCLUSION

Using previously reported data, the present study formally decomposed terminal silent-degradation judgments into d-prime and c. Larger reliability loss corresponded to greater discriminability, while high initial reliability combined greater discriminability with a more conservative relative reporting policy. Safe supervisory design should therefore improve both the diagnostic quality of reliability-change evidence and the calibration of the decision policy governing explicit intervention.

ACKNOWLEDGMENT

This study was supported by the National Natural Science Foundation of China (Grant Nos. 72571154, 72192822, and 72192824).

REFERENCES

Lee, J.D. and See, K.A. (2004) ‘Trust in automation: Designing for appropriate reliance’, Human Factors, 46(1), pp. 50–80. doi: 10.1518/hfes.46.1.50_30392.

Macmillan, N.A. and Creelman, C.D. (2005) Detection Theory: A User’s Guide. 2nd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Parasuraman, R. and Riley, V. (1997) 'Humans and automation: Use, misuse, disuse, abuse', Human Factors, 39(2), pp. 230–253. doi: 10.1518/001872097778543886.

Stanislaw, H. and Todorov, N. (1999) 'Calculation of signal detection theory measures', Behavior Research Methods, Instruments, & Computers, 31(1), pp. 137–149. doi: 10.3758/BF03207704.

Tan, C., Xu, X. and Ma, L. (2026a) 'Can humans timely and accurately perceive silent failures in automated decision aids? Effects of failure severity, initial reliability, and error type', International Journal of Industrial Ergonomics, 113, 103924. doi: 10.1016/j.ergon.2026.103924.

Tan, C., Xu, X., Wang, Y. and Ma, L. (2026b) 'Psychometric characterization of silent reliability degradation detection in automated decision aids', Human Error, Reliability, Resilience, and Performance, 213, pp. 15–25. doi: 10.54941/ahfe1007550.