

Estimating Cognitive Load via Process-based Telemetry for Adaptive Serious Games

Tobias Baumeister and Mark Oude Elberink

Friedrich-Alexander-Universität, Erlangen, Germany

tobias.baumeister@fau.de

mark.oude.elberink@fau.de

Abstract: Traditional cognitive load measures interrupt gameplay: the NASA-TLX requires time to complete, physiological sensors demand lab conditions. Neither integrates cleanly into a serious game session. This study tests whether sixteen process-based interaction metrics, captured automatically through in-game telemetry, can track perceived cognitive load without interrupting play. Thirty university students played LUTLAB (a puzzle-based serious game for learning FPGA hardware concepts) across ten levels. The game's telemetry logged mouse movement speed and distance, movement pause patterns (contemplation and hesitation frequencies, pause ratios), duplicate game states, test-button and tool usage, click frequency and burstiness, help-seeking behavior, and dwell time in task rooms. Players completed the Raw Task Load Index (Raw-TLX) after each level. Spearman rank correlations were computed between every metric and Raw-TLX scores across 150 level-observations. One metric held up: dwell time in task rooms correlated with perceived load at ($p = 0.48\text{--}0.66$, all $p < 0.01$) in every one of the five analyzed levels. Active mouse movement speed reached stronger correlations in individual levels (up to $p = -0.67$ in an extended analysis) but did not hold globally. Duplicate game states and contemplation pause frequency predicted load primarily in Level 3, the first level where players could freely interact with game elements. Most metrics showed weak or no correlation with Raw-TLX scores. What this means for adaptive game design is direct: there is no single behavioral signal that works uniformly across all task types. A metric's predictive value is tied to the structure of the specific task, not to any general property of the metric itself. Cognitive load indicators cannot simply be applied across an entire game; they must be selected and validated within the context of each module. This study provides systematic empirical evidence for this context-dependency across multiple levels of the same game, establishing a basis for task-aware behavioral sensing in adaptive serious games.

Keywords: Serious games, Cognitive load, Telemetry, Interaction metrics, NASA-TLX, Education

1. Introduction

Serious games pursue goals beyond entertainment, such as knowledge acquisition (Susi, Johannesson and Backlund, 2007). Maintaining intrinsic motivation and a state of flow (Csikszentmihalyi, 1990) requires balancing task difficulty with learner ability. Cognitive Load Theory (CLT) (Sweller, 1988) provides a framework for describing this balance: when the demands on working memory exceed its capacity, learning is impaired. Established measurement methods such as the NASA-TLX questionnaire (Hart and Staveland, 1988) or physiological sensors (Brunken, Plass and Leutner, 2003) are limited in gameplay contexts, as they can interrupt the game flow (Steinmaurer, Sackl and Gutl, 2021) or require specialized hardware.

A promising alternative is the use of behavioral data collected unobtrusively during gameplay. Prior work has shown that mouse movements change systematically under increased cognitive load (Grimes and Valacich, 2015; Arshad, Wang and Chen, 2013). Other interaction features, like click patterns, error rates, or repeated game states, are also potential indicators. While individual behavioral indicators have been linked to cognitive load (Sevcenko et al., 2021; Seyderhelm and Blackmore, 2023), a systematic comparison of multiple telemetry-based metrics within a single serious game is still lacking.

This matters because adaptive games can only scale difficulty in real time if they can sense load in real time. Unlike questionnaires or physiological sensors, telemetry is unobtrusive, needs no hardware, and works at classroom scale – making load-adaptive difficulty feasible outside the lab. Knowing which signals track load, and under what task conditions, is a precondition for such systems.

This paper addresses this gap through a systematic, exploratory comparison of 16 telemetry-based metrics as cognitive load indicators in the serious game LUTLAB.

2. Related Work

Grimes and Valacich (2015) found that average mouse speed decreases under increased cognitive load, while euclidean distance of individual mouse movements increases. Rheem, Verma and Becker (2018) confirmed this effect and distinguished it from purely motor difficulty. Arshad, Wang and Chen (2013) classified mouse movement pauses into hesitation (<1s) and contemplation (1–5s) pauses and found significant increases in

both pause types under higher cognitive load, though hesitation pauses were noted as harder to attribute directly to cognitive load.

Sevcenko et al. (2021) pursued a similar goal to ours: estimating cognitive load in a serious game using hidden metrics, employing Normalized Gaming Time (comparable to our dwell time metric) and temporal action density decay. Seyderhelm and Blackmore (2023) introduced a virtual detection response task (virDRT) for in-game cognitive load measurement, though in a different game context.

3. Method

3.1 Game Environment

LUTLAB is a puzzle-based serious game that teaches fundamentals of Field Programmable Gate Arrays (FPGA). Players traverse rooms within each level, configuring circuits (CLBs, switchboxes, cables) to implement specified logic functions – from simple gates to adders and ALUs. Levels contain three room types: interconnect rooms, circuit rooms, and quiz rooms. A test button allows players to verify their configurations at any time. The game comprises 10 levels of increasing complexity.

3.2 Metrics

We computed 16 metrics in seven categories from the telemetry data (Table 1). Metrics were either derived from the literature or defined as part of our exploratory investigation. Where applicable, metrics were normalized by dwell time. Dwell time itself was measured as the total time spent in task-relevant rooms per level.

Table 1: Overview of analyzed metrics by category.

Category	Metrics
Mouse movement	Speed (overall), speed (active only), distance
Movement pauses	Contemplation/Hesitation pause frequency, mouse/character pause ratio
State duplicates	Count (total, circuit-only, switchboxes-only), frequency
Button/tool usage	Test button frequency, tool switch frequency
Click patterns	Click frequency, click burstiness (per-room, per-level)
Help-seeking	Help button frequency
Dwell time	Total time in task rooms

Several metrics require clarification. *Mouse speed (active)* averages instantaneous speed only over time steps where the mouse was moving, whereas *mouse speed (overall)* divides total distance by total dwell time. *State duplicates* are specific to LUTLAB: each circuit configuration is represented as a hash, and a duplicate is counted whenever a previously seen configuration reoccurs, indicating that a player has unknowingly returned to an already-tried solution – a behavior consistent with working memory overload under CLT (Sweller, 1988). *Contemplation pauses* (1 – 5 s) and *hesitation pauses* (100 ms – 1 s) classify mouse standstill periods by duration following Arshad, Wang and Chen (2013). *Click burstiness* quantifies temporal regularity of clicks using the scalar metric by Goh and Barabási (2008) where $B = -1$ indicates perfectly periodic clicks and $B = 1$ indicates extreme burst-like patterns.

3.3 Study Design

The study was conducted during two exercise sessions of a Computer Architecture course at Friedrich-Alexander-Universität Erlangen-Nürnberg. Thirty students participated voluntarily, playing LUTLAB for up to 80 minutes. After completing each level, a modal Raw-TLX questionnaire (Hart, 2006) captured subjective cognitive load on six 0–100 scales. Data were collected anonymously via a telemetry interface. Only completed levels (with submitted questionnaires) were included in the analysis.

4. Results

The relationship between each metric and the Raw-TLX total score was quantified using Spearman rank correlations (ρ), classified following Dancey and Reidy (2007) as weak ($0.10 \leq |\rho| < 0.40$), moderate ($0.40 \leq |\rho| < 0.60$), or strong ($|\rho| \geq 0.60$).

Levels 0–4 were completed by all 30 participants; completion rates dropped monotonically from Level 5 onward. Because of this, the analysis was conducted with two groups: the full group (n=30, Levels 0–4) includes all participants but only the first five levels; the extended group (n=22, Levels 0–6) is restricted to those who completed Level 6. The extended group showed systematically lower dwell times and Raw-TLX scores already in Levels 0–4, indicating a selection effect: participants who reached Level 6 were faster and reported lower subjective cognitive load early on.

The per-level analysis revealed substantial variability (Fig. 1). Dwell time was the only metric consistently correlated across all five levels (p between 0.48 and 0.66, all $p < 0.01$). All other metrics showed level-dependent patterns:

- *Mouse speed (overall)* peaked in Level 2 ($p = -0.64$, $p < 0.001$) but was negligible in Levels 0–1.
- *State duplicates* concentrated their signal in Level 3 ($p = 0.58$, $p < 0.001$), the first non-tutorial level with free interaction, and showed weak-to-moderate correlations in Levels 1 and 4.
- *Contemplation pause frequency* was significant only in Level 3 ($p = 0.55$, $p < 0.01$).
- *Test button frequency* showed its strongest correlations in Levels 1–2 ($p \approx -0.50$, $p < 0.01$).

In the extended group, active mouse speed showed more consistent correlations across levels (p between -0.45 and -0.67), while state duplicate correlations weakened considerably – consistent with the selection effect producing a more homogeneous, lower-load subgroup.



Figure 1: Spearman correlations between metrics and Raw-TLX per level (full group, n=30). Dwell time is the only metric with consistent moderate-to-strong correlations across all levels

5. Discussion and Conclusion

The central finding of this study is the *context-dependence of interaction metrics* as cognitive load indicators. With the exception of dwell time, no metric showed consistent correlations with subjective cognitive load across game levels. The suitability of a metric depends on whether the task structure allows load-related differences to manifest in the measured behavior.

This effect is most visible in the contrast between Level 2 (a structured tutorial) and Level 3 (the first free-form level): Level 2 shows strong mouse speed correlations, while Level 3 reveals state duplicate and contemplation

pause signals that are absent in tutorial levels. The shift suggests that freer interaction contexts enable a wider range of behavioral indicators to become informative.

Dwell time emerged as the most robust indicator, consistent with Sevchenko et al.'s (2021) finding for Normalized Gaming Time. Notably, LUTLAB imposes no time limit, suggesting this relationship holds independently of time pressure. *Active mouse speed* showed promising consistency in the extended group, aligning with Grimes and Valacich (2015) and Rheem, Verma and Becker (2018). *State duplicates*, theoretically grounded in CLT's working memory limitations, were informative primarily in levels allowing free circuit configuration.

Several limitations qualify these findings. The sample ($n=30$) was small and self-selected from a single course, constraining statistical power in later levels and generalizability. How informative a metric appears is itself sensitive to sample composition: in the faster, lower-load subgroup that reached Level 6, state-duplicate correlations weakened while mouse-speed ones strengthened, so a different sample could have surfaced a different set of indicators. Our claim also generalizes only as a *principle*: because the study rests on LUTLAB, mechanic-specific metrics such as state duplicates are unlikely to transfer to all games, whereas the requirement to select and validate load indicators per task should. Dwell time and active mouse speed are the most plausible cross-game candidates, being mechanic-agnostic and consistent with prior work (Sevchenko et al., 2021; Grimes and Valacich, 2015), but this needs replication in other games. Dwell time is further confounded by non-task causes of long room times – a stuck mouse, a break, an interruption – which telemetry alone cannot separate from genuine effort; gating dwell time on concurrent activity (mouse, click, or state-change events) or other kinds of attention detection could disambiguate these. Finally, Raw-TLX was collected only per level, potentially masking room-type variation. Future work should use larger, more representative samples, finer-grained load measurement (e.g., per-room single-item scales, Paas (1992)), activity-gated dwell time, and cross-game replication.

Ethics Declaration: An Ethical clearance was not required for this research. Participants provided informed consent. Those who declined received a version of the game without telemetry.

AI Declaration: AI was partially used to summarize, improve structure and language of this paper.

References

- Arshad, S., Wang, Y. and Chen, F. (2013) "Analysing mouse activity for cognitive load detection", Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration. OzCHI '13: Augmentation, Application, Innovation, Collaboration, Adelaide Australia: ACM, pp. 115–118. doi:10.1145/2541016.2541083.
- Brunken, R., Plass, J.L. and Leutner, D. (2003) "Direct Measurement of Cognitive Load in Multimedia Learning", *Educational Psychologist*, 38(1), pp. 53–61. doi:10.1207/S15326985EP3801_7.
- Csikszentmihalyi, M. (1990) *Flow: the psychology of optimal experience*. 1st edn. New York: Harper & Row.
- Dancey, C.P. and Reidy, J. (2007) *Statistics without maths for psychology*. 4th edn. Harlow: Pearson Prentice Hall.
- Goh, K.-I. and Barabási, A.-L. (2008) "Burstiness and memory in complex systems", *EPL (Europhysics Letters)*, 81(4). doi:10.1209/0295-5075/81/48002.
- Grimes, G.M. and Valacich, J.S. (2015) "Mind Over Mouse: The Effect of Cognitive Load on Mouse Movement Behavior", ICIS 2015 Proceedings, p. 9. Available at: <https://aisel.aisnet.org/icis2015/proceedings/HCI/9/> (Accessed: 17 January 2026).
- Hart, S.G. (2006) "Nasa-Task Load Index (NASA-TLX); 20 Years Later", Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50(9), pp. 904–908. doi:10.1177/154193120605000909.
- Hart, S.G. and Staveland, L.E. (1988) "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research", in P.A. Hancock and N. Meshkati (eds) *Human Mental Workload*. North-Holland (Advances in Psychology), pp. 139–183. doi:10.1016/S0166-4115(08)62386-9.
- Paas, F.G.W.C. (1992) "Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach", *Journal of Educational Psychology*, 84(4), pp. 429–434. doi:10.1037/0022-0663.84.4.429.
- Rheem, H., Verma, V. and Becker, D.V. (2018) "Use of Mouse-tracking Method to Measure Cognitive Load", Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 62(1), pp. 1982–1986. doi:10.1177/1541931218621449.
- Sevchenko, N. et al. (2021) "Measuring Cognitive Load Using In-Game Metrics of a Serious Simulation Game", *Frontiers in Psychology*, 12, p. 572437. doi:10.3389/fpsyg.2021.572437.
- Seyderhelm, A.J.A. and Blackmore, K.L. (2023) "How Hard Is It Really? Assessing Game-Task Difficulty Through Real-Time Measures of Performance and Cognitive Load", *Simulation & Gaming*, 54(3), pp. 294–321. doi:10.1177/10468781231169910.
- Steinmaurer, A., Sackl, M. and Gutl, C. (2021) "Engagement in In-Game Questionnaires - Perspectives from Users and Experts", 2021 7th International Conference of the Immersive Learning Research Network (iLRN). 2021 7th

- International Conference of the Immersive Learning Research Network (iLRN), Eureka, CA, USA: IEEE, pp. 1–7.
doi:10.23919/iLRN52045.2021.9459373.
- Susi, T., Johannesson, M. and Backlund, P. (2007) Serious Games – An Overview. HS-IKI-TR-07-001. University of Skövde.
Available at: <https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A2416&dswid=-2433> (Accessed: 15 March 2026).
- Sweller, J. (1988) “Cognitive Load During Problem Solving: Effects on Learning”, *Cognitive Science*, 12(2), pp. 257–285.
doi:10.1207/s15516709cog1202_4.