SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper

Analyzing Attention Patterns in Children with ADHD through Interaction Data from Educational Games

Dau Minh Hai[1], Ms. Hoang Thi Hong Minh[2]

[1]Hanoi - Amsterdam High School for the Gifted, Thach Ha, Ha Tinh E-mail: dauhai0409@gmail.com
[2]Mokpo National University – Korea Group 16, Dai Mo Ward, Hanoi. E-mail: hoanghongminh13@gmail.com

Abstract
Interaction data generated by serious games provide an opportunity to characterize cognitive behavior at a substantially finer temporal resolution than conventional session-level performance scores. However, game-based ADHD research has largely focused on aggregate outcomes or participant classification, while the sequential organization and within-task dynamics of attention-related behavior remain comparatively underexplored. This study investigates attention-related behavioral patterns derived from Attention Slackline and Attention Robots in the BALLADEER dataset. Among 85 children aged 6-12 years, the primary comparison cohort comprised 58 participants with diagnosed=yes and 18 with diagnosed=no; nine participants with diagnosed=undetermined were excluded from primary group comparisons. Across 254 Attention Slackline sessions, 4,820 of 4,822 expected target opportunities were successfully reconstructed. Mixed-effects models were used to analyze omission, commission, and reaction time (RT), while a Hidden Semi-Markov Model (HSMM) was used to infer latent behavioral states from sequential interaction data. Two machine-learning models were evaluated using nested five-fold participant-level cross-validation, and cross-game consistency was examined in 70 participants with sufficient data from both tasks.
The ADHD group showed higher odds of omission than the comparison group (OR=2.51; 95% interval: 2.25-2.80) and higher odds of commission (OR=1.83; 1.60-2.09), whereas no reliable ADHD-related difference in RT or ADHD-specific increase in omission over time was observed. The HSMM identified three latent behavioral states. An omission-dominant state showed higher adjusted occupancy in the ADHD group (+7.0 percentage points; p=0.0325), although the effect did not remain significant after false discovery rate correction (q=0.068). Out-of-fold AUROC ranged from 0.53 to 0.59, and temporal features did not improve discrimination relative to conventional behavioral features. Cross-game associations between Attention Slackline and Attention Robots were generally weak.
These findings indicate that game interaction data can reveal attention-related behavioral differences at both event and latent-state levels, but do not currently support accurate ADHD classification or strong generalization across tasks. The primary contribution therefore lies in behavioral computing and temporal behavioral modeling rather than automated ADHD diagnosis.

Keywords: ADHD; attention patterns; serious games; interaction data; behavioral computing; Hidden Semi-Markov Model; game learning analytics.

1. Introduction

Attention-deficit/hyperactivity disorder (ADHD) is a neurodevelopmental disorder characterized by symptoms of inattention and/or hyperactivity-impulsivity. In attention-related assessment tasks, omission errors, commission errors, reaction time (RT), and reaction-time variability (RTV) are commonly used to characterize behavioral performance. Meta-analytic evidence indicates that children with ADHD tend to exhibit more errors and greater response variability than typically developing children, although these measures are insufficiently specific to serve as stand-alone diagnostic indicators [1], [2], [15]-[17].
The increasing use of **game-based cognitive assessment** provides a substantially richer source of behavioral data than conventional end-point testing. Rather than recording only final performance scores, digital games can capture response sequences, stimulus timing, errors, RT, and visual interaction behavior. Systematic reviews have shown growing use of digital games for attention and ADHD assessment, while also highlighting substantial heterogeneity in task design, gameplay characteristics, and validation procedures [3], [4], [18].
From a computer-science perspective, such interaction logs naturally fall within **game learning analytics** and **behavioral computing**, where fine-grained telemetry is transformed into computational representations of user behavior [8]. The analytical problem is therefore not limited to determining how well a participant performs overall, but extends to reconstructing, representing, and modeling how behavioral patterns evolve during interaction.
Attention Slackline is a serious game developed to assess attention-related performance in ADHD. Previous studies have reported associations between game performance and established attention measures and have demonstrated some ability to discriminate participants with and without ADHD, particularly among younger children [5], [6]. However, these studies have primarily relied on **aggregate performance**

measures. Such aggregation may conceal substantial within-session structure: two participants with similar overall error rates may nevertheless differ in when errors occur, whether errors cluster temporally, or how they transition between behavioral modes.

Sequential modeling provides a computational framework for exploiting this structure. Hidden Semi-Markov Models (HSMMs) extend conventional Hidden Markov Models (HMMs) by explicitly modeling state duration, making them particularly suitable for behavioral processes that may persist across multiple interaction events before transitioning [7]. In parallel, machine-learning approaches have been applied to ADHD classification using gameplay-derived features [9]. However, in small datasets, feature engineering or model tuning performed outside the validation procedure can lead to data leakage and optimistically biased performance estimates [12]-[14].

The present study uses BALLADEER, a multimodal dataset containing game-based tasks, eye-tracking data, EEG, and physiological signals collected from children and adolescents [10]. Because the fixed title of this study uses the term educational games, this term is operationalized here in the broader sense of **game-based cognitive assessment environments**, rather than games designed primarily for academic instruction. The present analysis focuses exclusively on interaction data from Attention Slackline and Attention Robots.

Three research questions are addressed:

RQ1: How do attention-related interaction patterns differ between children with ADHD and the comparison group across time-on-task and difficulty levels in Attention Slackline?

RQ2: What latent attention-related behavioral states can be inferred from sequential interaction data, and how do their occupancy, duration, and transition dynamics differ between groups?

RQ3: To what extent are the attention-related patterns identified in Attention Slackline consistent with block-level performance and visual-search behavior in Attention Robots?

The main computational contributions are fourfold:

(1) reconstruction of event-level behavioral streams by combining stimulus schedules with interaction logs;

(2) temporal representation of attention-related behavior at the event and participant levels;

(3) unsupervised latent-state inference using HSMMs;

(4) evaluation of representation generalizability through participant-level cross-validation and cross-game analysis.

Accordingly, the study is positioned as a **computational behavioral modeling** investigation rather than an attempt to develop an automated diagnostic system for ADHD.

2. Related Work

2.1. Game-Based Assessment of Attention

Game-based assessment has evolved through several approaches, including gamification of conventional cognitive tasks, development of novel games targeting specific cognitive constructs, and analysis of behavioral data naturally generated during gameplay [3]. However, game mechanics such as feedback, stimulus rate, visual presentation, and reward structures may themselves influence behavior. Consequently, game performance cannot automatically be interpreted as a neutral measurement of attention [3], [18].

In ADHD research, systematic reviews suggest that digital games may provide useful complementary assessment information, although findings remain heterogeneous across tasks, behavioral features, and validation methodologies [4]. Similarly, systematic reviews of Continuous Performance Tests (CPTs) indicate that omission, commission, and RT measures have descriptive and discriminative value, but their clinical utility as independent ADHD identification tools remains limited [17], [19].

Attention Slackline belongs to this class of game-based assessment environments. Earlier work demonstrated that the task captures attention-related behavioral differences [5], while a subsequent validation study reported discrimination between ADHD and non-ADHD participants, with stronger effects among younger children [6]. The present study extends these previous investigations by moving from session-level performance measures to event-level and sequential behavioral representations.

2.2. Temporal and Sequential Behavioral Modeling

RTV is commonly interpreted as a measure of **intra-individual variability** and is consistently elevated in ADHD across a variety of cognitive tasks [1], [15]. However, RTV is not specific to ADHD and has also been observed across other neurodevelopmental conditions [16]. Evidence regarding ADHD-specific time-on-task deterioration is likewise inconsistent, suggesting that temporal group differences should be estimated directly rather than assumed a priori.

Aggregate statistics cannot directly represent transitions between different behavioral modes. HSMMs provide an alternative sequential representation in which latent states are characterized by both emission distributions and explicit duration models [7]. Such models are therefore well suited to investigating whether errors and RT patterns organize into distinct behavioral regimes during gameplay.

2.3. Predictive Evaluation and Cross-Task Analysis

Gameplay-derived features have previously been used in machine-learning-based ADHD assessment [9]. However, predictive performance is interpretable only when preprocessing, feature selection, and hyperparameter tuning are confined to the training data within the validation procedure [12]-[14]. The present study therefore uses nested participant-level cross-validation and treats classification as a secondary exploratory analysis rather than the primary contribution.

The main methodological gap addressed here is the integration of **event reconstruction, temporal behavioral modeling, latent-state inference, participant-level predictive validation, and cross-game comparison** within a single analytical framework.

3. Methodology

3.1. Dataset and Cohort

This study is a retrospective secondary observational analysis of the BALLADEER dataset [10]. Reporting of the observational component was informed by STROBE [11], while the predictive analysis followed principles of participant-level out-of-sample evaluation and leakage prevention emphasized in TRIPOD+AI and related methodological work [12]-[14].

The demographic file used in the analysis contained 158 participant records, including 85 children aged 6-12 years. Within this age range, 58 participants had diagnosed=yes, 18 had diagnosed=no, and nine had diagnosed=undetermined. Primary group comparisons therefore included 76 participants: 58 in the ADHD group and 18 in the comparison group. Participants with undetermined diagnostic status were excluded from the primary inferential and predictive analyses.

The ADHD group had a mean age of 9.71±1.74 years and included 37 male and 21 female participants. The comparison group had a mean age of 8.67±1.57 years and included 13 male and five female participants.

Because the dataset variables group and diagnosed were not fully concordant, sensitivity analyses additionally considered alternative cohort definitions, including participants aged 6-11 years, grouping according to group, a concordant-only subset, and a subset with complete Slackline sessions.

3.2. Event Reconstruction and Temporal Alignment

The age-filtered cohort contained 254 Attention Slackline TAGS sessions: 85 at Level 1, 84 at Level 6, and 85 at Level 11. Attention Robots included 82 GAME_DATA files and 79 EYE_TRACKING_DATA files.

Slackline TAGS logs record presses and missed required responses but do not provide the complete sequence of presented stimuli. Correctly ignored non-target stimuli therefore leave no corresponding log entry. To recover the event stream, interaction logs were combined with the original Slackline stimulus schedule.

Target definitions differed across levels. At Level 1, the target was a circle; at Level 6, either a circle or a square was a target; at Level 11, the target was the circle→square sequence, with the target opportunity assigned at sequence completion.

Because generalTime and scheduled stimulus times used different temporal origins across sessions, event timing was aligned separately for each session. For valid response j in session s, the observed stimulus onset was estimated as

$$t_{sj}^{obs} = g_{sj} - RT_{sj}, (1)$$

where g_{sj} is generalTime and RT_{sj} is the recorded reaction time.

The session-specific temporal offset was estimated as

$$\Delta_s = \mathrm{median}_{j\in\mathcal{K}_s}\left(t_{sj}^{obs} - t_{sj}^{flag}\right), (2)$$

where $\mathcal{K}_s$ denotes the set of valid responses used for alignment and t_{sj}^{flag} denotes scheduled stimulus time.

Events were subsequently matched using stimulus order, stimulus type, and temporal proximity. Dynamic sequence alignment was used when incomplete event sequences required preservation of ordering constraints.

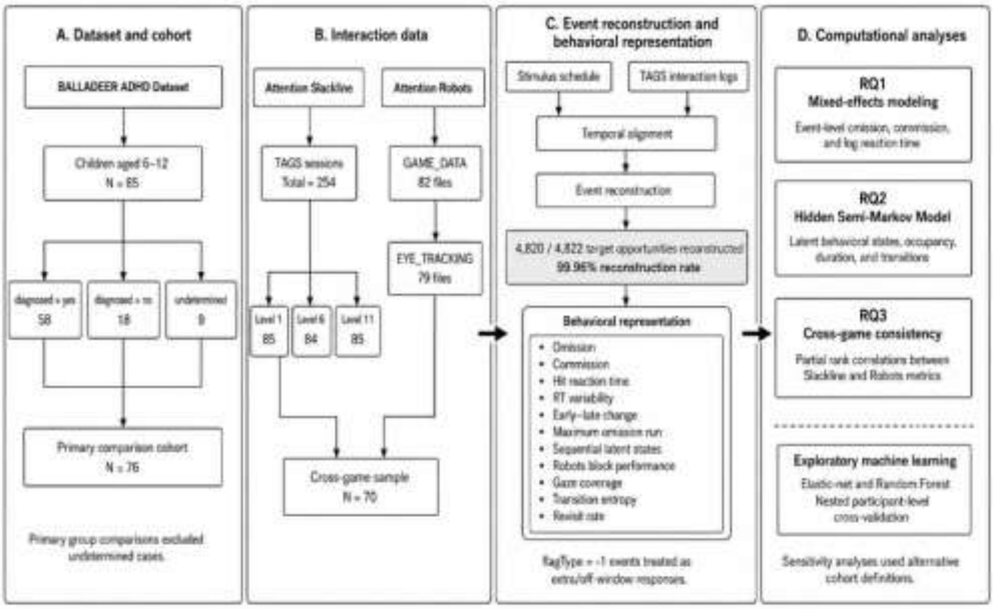
Target opportunities were coded as hits when reacted=True and omissions when reacted=False. For omissions, reactionTime=-1 was converted to missing and was not treated as a numerical RT. Nonduplicated incorrect responses associated with a concrete flag type were mapped to non-target opportunities. Responses with flagType=-1 were treated as extra or off-window responses and excluded from the primary commission-rate calculation.

To represent time-on-task consistently across sessions, event time was normalized as

$$\tau_{sj} = \frac{t_{sj} - t_{s,\text{start}}}{t_{s,\text{end}} - t_{s,\text{start}}}, \qquad 0 \le \tau_{sj} \le 1, (3)$$

Relative time was subsequently centered at 0.5 for inferential modeling.

Figure 1. Study cohort, event-reconstruction pipeline, behavioral representation, and computational analyses.



3.3. Behavioral Representation

Attention Slackline behavior was represented using omission rate, commission rate, median hit RT, median absolute deviation of hit RT, late-minus-early changes in omission, commission, and log RT, and maximum consecutive omission run.

Attention Robots contributed block-level work speed, omission, commission, and correct-response measures. Eye-tracking features included robot-observation coverage, number of unique observed cells, transition rate, revisit rate, and normalized transition entropy.

Eye-tracking gaps were treated as unobserved intervals rather than as evidence of inattention.

3.4. Event-Level Mixed-Effects Modeling

RQ1 was evaluated using event-level mixed-effects models. Omission and commission were modeled as binary outcomes with participant-specific random intercepts. Fixed effects included ADHD status, centered relative time, difficulty level, ADHD×Time, ADHD×Level, age, and sex:

$$\operatorname{logit}\left[P\left(Y_{isj}=1\right)\right]=\beta_0+\beta_1 ADHD_i+\beta_2\tau^c_{isj}+\beta_3^\top Level_{isj}+\beta_4\left(ADHD_i\times\tau^c_{isj}\right)+\beta_5^\top\left(ADHD_i\times Level_{isj}\right)+\beta_6 Age_i+\beta_7 Sex_i+u_i,\ (4)$$

where Y_{isj} represents the binary behavioral outcome, τ^c denotes relative time centered at 0.5, and u_i denotes a participant-specific random intercept. The ADHD main effect therefore corresponds to the estimated group difference at the midpoint of Level 1.

RT for valid hits was log-transformed and analyzed using a linear mixed-effects model with the corresponding fixed effects and participant-level random effects. Benjamini-Hochberg false discovery rate (FDR) correction was applied within inferential families.

3.5. Sequential Latent-State Modeling

RQ2 was examined using a Hidden Semi-Markov Model (HSMM) fitted to target-event sequences from all 85 age-eligible participants with usable Slackline data. ADHD labels were not used during state estimation, ensuring that latent-state learning remained unsupervised. Subsequent group comparisons were restricted to participants with determined diagnostic status.

Each event was represented by omission status and log RT when a valid response was available. The state-specific emission model was

$$p(o_t, r_t \mid z_t = k) = \mathrm{Bernoulli}(o_t; \pi_k)\left[\mathcal{N}\left(\log r_t; \mu_k, \sigma_k^2\right)\right]^{1-o_t},\ (5)$$

where z_t denotes latent state k, ot denotes omission status, and r_t denotes RT.

State duration was modeled using a shifted Poisson distribution. Candidate models with two to five states were compared using participant-level held-out log-likelihood. Final model selection considered both out-of-sample fit and interpretability, particularly avoiding solutions containing extremely small states.

For the selected model, state occupancy, duration, and transition characteristics were summarized at the participant level. ADHD-related group differences were adjusted for age and sex and corrected for multiple testing using FDR.

3.6. Participant-Level Predictive Evaluation

A secondary exploratory analysis evaluated whether the derived behavioral features supported participant-level ADHD discrimination.

Two classifiers were examined: elastic-net logistic regression and Random Forest. The conventional feature set included age, sex, omission rate, commission rate, and median hit RT. The temporal feature set additionally included RT variability, late-minus-early behavioral changes, and maximum omission run.
Model evaluation used nested five-fold participant-level cross-validation. Imputation, standardization, and hyperparameter tuning were performed exclusively within the training folds to prevent information leakage.
Elastic-net used l1_ratio=0.5, with C selected from 0.1, 1, and 10 using inner three-fold cross-validation. Random Forest used 200 trees, with max_depth selected from 2, 4, or unrestricted depth and min_samples_leaf selected from 3 or 5.
The primary performance measure was pooled out-of-fold AUROC; balanced accuracy and Brier score were reported as complementary metrics.

3.7. Cross-Game and Sensitivity Analyses

RQ3 examined consistency between Attention Slackline and Attention Robots using partial rank correlations adjusted for age, sex, and ADHD status. Analyses compared overall performance, within-task deterioration, and selected gaze-derived measures. FDR correction was applied within the corresponding correlation families.
Sensitivity analyses repeated the main group effects and exploratory classification analysis under alternative cohort definitions. Eye-tracking associations were additionally evaluated under different minimum observation-coverage thresholds.

4. Results

4.1. Cohort and Reconstruction Quality

The age-filtered dataset included 85 children aged 6-12 years, with 76 participants included in the primary ADHD-versus-comparison analyses.
Across 254 Attention Slackline sessions, 4,822 target opportunities were expected from the stimulus schedules. Of these, 4,820 were successfully reconstructed, corresponding to a reconstruction rate of **99.96%**. The two remaining opportunities were retained as unobserved rather than being coded as omissions.
Within the primary cohort, Level 1, Level 6, and Level 11 data were available for 76, 75, and 76 participants, respectively. Attention Robots GAME_DATA were available for 73 participants. Seventy participants had sufficient Slackline and Robots data for the cross-game analysis.
Participant characteristics, data availability, and descriptive behavioral measures are summarized in Table 1.

Table 1. Participant characteristics, data availability, and descriptive behavioral measures.

Characteristic	**ADHD**	**Comparison**
Participants, n	58	18
Age, years, mean ± SD	9.71 ± 1.74	8.67 ± 1.57
Male / Female, n	37 / 21	13 / 5
Slackline Level 1 available, n	58	18
Slackline Level 6 available, n	58	17
Slackline Level 11 available, n	58	18
Robots GAME_DATA available, n	56	17
Robots eye tracking available, n	54	16
Omission rate, mean	12.65%	9.27%
Commission rate, mean	4.57%	4.12%
Median hit RT, mean, s	1.001	0.994

MAD hit RT, mean, s	0.148	0.127
Maximum omission run, mean	2.19	1.83

4.2. RQ1: Group and Temporal Effects

At the participant level, mean omission rates were 12.65% in the ADHD group and 9.27% in the comparison group. Mean commission rates were 4.57% and 4.12%, respectively, whereas median hit RT was similar between groups.

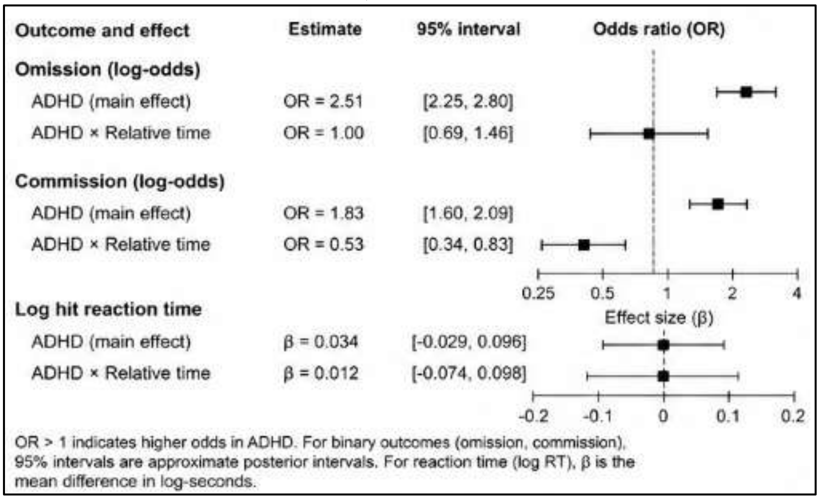
The omission model included 4,307 target opportunities. At the midpoint of Level 1, ADHD was associated with substantially higher odds of omission (OR=2.51, 95% interval 2.25-2.80). Although omission increased over time in the reference group, the ADHD×Time interaction was approximately null, indicating no evidence of stronger omission deterioration over time in the ADHD group.

The commission model included 7,192 non-target opportunities. ADHD was associated with higher odds of commission at the reference condition (OR=1.83, 95% interval 1.60-2.09). The ADHD×Time interaction was below unity in the primary model, but its direction and magnitude were not fully stable under alternative cohort definitions. This interaction was therefore not treated as a robust ADHD-specific temporal effect.

The RT model included 3,793 valid hits. Neither the ADHD main effect nor the ADHD×Time interaction provided evidence of a reliable group difference in RT.

Overall, RQ1 was partially supported: the clearest group differences concerned omission and commission rather than mean RT or ADHD-specific time-on-task deterioration.

Figure 2A. Event-level ADHD group effects and ADHD-by-time interactions for omission, commission, and reaction time.



4.3. RQ2: Latent Behavioral States

Participant-level held-out log-likelihood improved from the two-state to the five-state models. However, the four- and five-state solutions introduced an extremely small state representing approximately 1.4% of events. The three-state model was therefore selected as the largest solution that retained an interpretable state structure without introducing a very small residual state.

The selected HSMM identified three behavioral states. **S1 (stable-fast)** accounted for 47.61% of decoded events, with an omission rate of 0.04%, median RT of 0.901 s, and mean modeled duration of 7.50 trials. **S2 (stable-slower)** accounted for 40.00% of events, with an omission rate of 0.05%, median RT of 1.167 s, and mean duration of 3.18 trials. **S3 (omission-dominant)** accounted for 12.39% of events and was characterized by a 91.62% omission rate and mean duration of 1.47 trials.

After adjustment for age and sex, S3 occupancy was 7.00 percentage points higher in the ADHD group (95% CI 0.58-13.43; p=0.0325). However, this result did not remain significant after FDR correction (q=0.068). S1 occupancy was 12.64 percentage points lower in the ADHD group and also yielded q=0.068. No state-duration or transition-dynamic differences remained significant after FDR correction.

Accordingly, RQ2 provides suggestive, but not confirmatory, evidence that ADHD is associated with greater occupancy of an omission-dominant behavioral state.

Figure 2B. Behavioral profiles of the three HSMM states and adjusted ADHD effects on state occupancy.

	S1 Stable-fast	S2 Stable-slower	S3 Omission-dominant
Decoded event fraction	47.61%	40.00%	12.39%
Observed omission rate	0.04%	0.05%	91.62%
Median hit RT (s)	0.901	1.167	0.200*
Mean duration (trials)	7.50	3.18	1.47
Adjusted ADHD occupancy effect (percentage points)	-12.64	+5.64	**+7.00**
p-value	0.0453	0.2204	**0.0325**
FDR q-value	0.0680	0.2204	**0.0680**

State names reflect behavioral characteristics inferred from emissions. * Not interpretable
RT in S3 is not interpretable due to the high omission rate.
Occupancy effects were adjusted for age and sex.

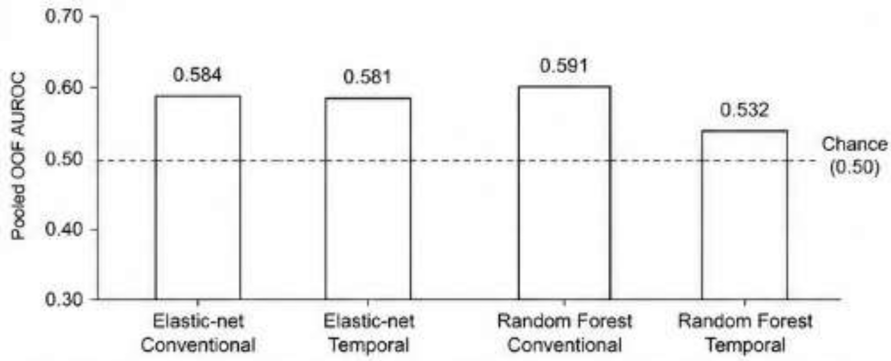
4.4. Exploratory Participant-Level Prediction

Participant-level prediction was weak across both algorithms and feature sets. Elastic-net logistic regression achieved pooled out-of-fold AUROCs of 0.584 using conventional features and 0.581 using temporal features. Random Forest achieved corresponding AUROCs of 0.591 and 0.532.

Thus, adding temporal behavioral features did not improve out-of-sample discrimination in either model. The detailed AUROC, balanced-accuracy, and Brier-score results are reported in Table 2.

These findings indicate that the interaction-derived representations contain group-level behavioral information but do not provide sufficient separation for reliable individual-level ADHD classification in the present sample.

Figure 2C. Pooled out-of-fold participant-level classification performance for conventional and temporal behavioral feature sets.



	Elastic-net Conventional	Elastic-net Temporal	Random Forest Conventional	Random Forest Temporal
Balanced accuracy	0.552	0.580	0.555	0.559
Brier score	0.247	0.235	0.205	0.205

Temporal features did not improve pooled out-of-fold AUROC compared with conventional features in either model.

4.5. RQ3: Cross-Game Consistency

The cross-game analysis included 70 participants, comprising 54 ADHD and 16 comparison participants.

After adjustment for age, sex, and ADHD status, associations between conceptually related Slackline and Robots measures were generally weak. Overall omission showed little association across tasks (partial r=0.087), as did overall commission (partial r=0.019).

The largest observed association was between Slackline omission and Robots observation coverage (partial r=−0.230, p=0.0549, q=0.165), but this relationship did not survive FDR correction. No cross-game association reached the corrected significance threshold.

Sensitivity analyses showed greater robustness for group-level behavioral effects than for predictive performance. Across alternative cohort definitions, omission ORs remained approximately 2.31-3.15 and commission ORs remained approximately 1.79-2.52. ADHD-related S3 occupancy differences also retained the same positive direction, ranging from approximately +6.8 to +10.2 percentage points.

In contrast, temporal elastic-net AUROC varied from 0.413 to 0.662 across cohort definitions, indicating substantial sensitivity of participant-level prediction to label and cohort specification.
Overall, RQ3 did not support strong cross-game consistency. Attention Slackline and Attention Robots appear to capture partially distinct aspects of attention-related behavior rather than interchangeable measurements of a single construct.

Figure 2D. Partial rank correlations between Attention Slackline behavioral measures and Attention Robots performance and gaze measures.

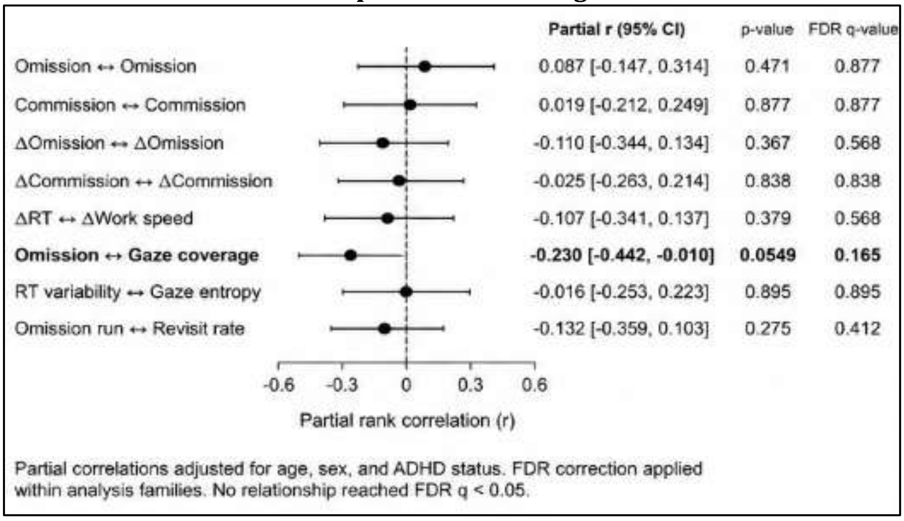


Table 2. Summary of the main inferential, latent-state, predictive, and cross-game results.

A. RQ1 — Event-level models

Analysis	Effect	Estimate	95% interval	p/q
Omission	ADHD	OR 2.511	2.251-2.802	—
Omission	ADHD × Time	OR 1.004	0.693-1.455	—
Commission	ADHD	OR 1.830	1.602-2.091	—
Commission	ADHD × Time	OR 0.529	0.338-0.827	—
Log RT	ADHD	β=0.034	−0.029-0.096	p=0.289
Log RT	ADHD × Time	β=0.012	−0.074-0.098	p=0.778

B. RQ2 — HSMM group effects

State	Adjusted ADHD occupancy effect	p	FDR q
S1 Stable-fast	−12.64 pp	0.0453	0.0680
S2 Stable-slower	+5.64 pp	0.2204	0.2204

S3 Omission-dominant	+7.00 pp	0.0325	0.0680

C. Exploratory machine learning

Model	Features	Pooled OOF AUROC	Balanced accuracy
Elastic-net	Conventional	0.584	0.552
Elastic-net	Temporal	0.581	0.580
Random Forest	Conventional	0.591	0.555
Random Forest	Temporal	0.532	0.559

D. RQ3 — Cross-game analysis

Relationship	Partial r	p	FDR q
Omission ↔ Omission	0.087	0.471	0.877
Commission ↔ Commission	0.019	0.877	0.877
ΔOmission ↔ ΔOmission	−0.110	0.367	0.568
ΔRT ↔ ΔWork speed	−0.107	0.379	0.568
Omission ↔ Gaze coverage	−0.230	0.0549	0.165

5. Discussion

The present study demonstrates that interaction logs from game-based cognitive tasks can be transformed into event-level and sequential behavioral representations, but that their descriptive value does not automatically translate into strong participant-level ADHD classification. Three findings are particularly important: omission and commission were consistently elevated in the ADHD group; the HSMM identified an omission-dominant state that tended to show higher occupancy among children with ADHD; and both predictive performance and cross-game correspondence remained limited.

5.1. Attention-Related Behavioral Differences

ADHD was associated with higher odds of both omission and commission after accounting for age, sex, difficulty level, time-on-task, and repeated observations. The direction of these effects was preserved across several sensitivity analyses, making them the most robust findings of the study.

These results are consistent with CPT evidence showing elevated omission and commission errors in ADHD [2], as well as with earlier Attention Slackline studies reporting ADHD-related performance differences [5], [6]. Nevertheless, reviews of CPT clinical utility emphasize that objective behavioral measures do not provide sufficient specificity for stand-alone ADHD diagnosis [17], [19]. Accordingly, omission and commission in the present study are better interpreted as digital behavioral measures of attention-related performance and response control rather than ADHD biomarkers.

In contrast, neither mean RT nor ADHD-specific time-on-task deterioration was reliably supported. This finding is compatible with evidence suggesting that ADHD-related performance differences may be more strongly reflected in response variability or distributional characteristics than in mean RT alone [1], [15], [16].

The results also argue against assuming a universal linear deterioration trajectory in ADHD. Behavioral performance is likely to depend on task structure, stimulus rate, difficulty, and other interaction characteristics, supporting direct modeling of group-by-time interactions rather than imposing a predefined decline pattern.

5.2. Latent Behavioral Representation

The HSMM provided a complementary representation of interaction sequences by decomposing behavior into three latent regimes rather than summarizing each session with a small number of aggregate measures.

Two states showed near-zero omission but differed substantially in RT, whereas the third state was overwhelmingly characterized by omission. Although ADHD participants showed higher estimated occupancy of the omission-dominant state, the FDR-corrected result q=0.068 supports only a cautious, suggestive interpretation.

Importantly, S3 should not be labeled an inattentive state in a clinical or cognitive sense. It is a latent behavioral state inferred from the observed emission structure. Without an independent measure of momentary cognitive state, a stronger psychological interpretation would exceed the evidence available in the dataset.

Methodologically, the HSMM is valuable because it explicitly represents state duration and sequence structure [7]. However, no reliable group differences in duration or transition dynamics were detected after correction for multiple comparisons. The present evidence therefore supports the value of HSMM as a behavioral representation framework, but not the existence of an ADHD-specific transition mechanism.

5.3. Limited Predictive Utility

Out-of-fold AUROC values of approximately 0.53-0.59 indicate that the evaluated interaction-derived features were insufficient for reliable individual-level ADHD classification. Furthermore, adding temporal features did not improve AUROC relative to the conventional feature set.

This does not contradict the statistically detectable group-level effects for omission and commission. A feature may show a meaningful population-level difference while still exhibiting substantial distributional overlap between individual participants, thereby limiting its predictive utility.

Sample size and class imbalance further constrained model stability, particularly because the primary comparison group contained only 18 diagnosed=no participants. The wide variation in AUROC under alternative cohort definitions additionally demonstrates substantial label sensitivity.

Participant-level nested cross-validation reduces optimistic evaluation bias, but it cannot eliminate uncertainty resulting from small sample sizes, imbalanced groups, or imperfect outcome definitions [13], [14]. Machine learning should therefore be interpreted in this study as a test of behavioral-feature generalizability rather than as evidence for a diagnostic system.

5.4. Cross-Game Consistency

Associations between Attention Slackline and Attention Robots were generally close to zero. This does not necessarily imply that either task lacks behavioral validity, because the two tasks differ in their interaction requirements.

Attention Slackline emphasizes sequential stimulus recognition and response control, whereas Attention Robots combines visual search with block-level task performance. Reviews of game-based assessment similarly emphasize that task mechanics and design features can influence the behavioral constructs expressed during interaction [3], [18].

The weak cross-game associations therefore suggest that attention-related behavior is multidimensional and task dependent. The negative relationship between Slackline omission and Robots observation coverage was the strongest observed association, but it did not survive FDR correction and should therefore be considered hypothesis-generating rather than co**nfirmatory.**

5.5. Strengths, Limitations, and Computer-Science Implications

The main technical strength of the study is the reconstruction of 99.96% of expected target opportunities from raw interaction logs, enabling the transition from session-level scores to event-level behavioral representations.

The analytical pipeline integrates temporal alignment, repeated-measures modeling, latent-state modeling, participant-level validation, and cross-task analysis. This structure is closely aligned with game analytics and behavioral computing, where the primary value of interaction telemetry lies in computational modeling rather than final task scores alone [8].

Several limitations should be considered. First, the study is a secondary analysis and therefore had no direct control over the original data-collection protocol. Second, the group and diagnosed variables were not fully concordant, introducing label uncertainty. Third, the comparison group was substantially smaller than the ADHD group. Fourth, Attention Robots and eye-tracking data were not complete for all participants, and eye-tracking gaps could only be treated as unobserved rather than interpreted behaviorally. Fifth, selection of the three-state HSMM reflected a balance between held-out likelihood and interpretability; higher-state solutions achieved better numerical likelihood but introduced extremely small states. Finally, the current study does not include external validation on an independent dataset.

From a computer-science perspective, the findings support the importance of storing fine-grained event telemetry, timestamps, reaction metadata, and gaze events in future game-based cognitive assessment systems rather than relying exclusively on aggregate scores. Such data enable sequence modeling, state-space approaches, and personalized behavioral analytics.

At the same time, the results demonstrate that richer behavioral representations do not necessarily translate into stronger predictive performance. The utility of increasingly complex representations should therefore be evaluated through out-of-sample generalization rather than inferred from within-dataset statistical differences alone.

6. Conclusion

This study developed a behavioral-computing pipeline for analyzing attention-related interaction patterns in children with ADHD using game-based cognitive tasks. Event reconstruction enabled Attention Slackline to be analyzed at stimulus level, while mixed-effects models showed that omission and commission were consistently elevated in the ADHD group. In contrast, no reliable evidence was found for ADHD-related differences in mean RT or stronger time-on-task deterioration in omission.

The HSMM identified three interpretable behavioral states, including an omission-dominant state that tended to show greater occupancy among children with ADHD, although this effect did not remain significant after FDR correction. Temporal features did not improve participant-level classification performance, and cross-game consistency between Attention Slackline and Attention Robots was weak.

The principal contribution of this work therefore lies in the reconstruction, representation, and computational modeling of sequential behavioral interaction data, rather than automated ADHD diagnosis. Future work should evaluate the stability and generalizability of these behavioral representations using independent datasets, larger comparison cohorts, and multiple interaction tasks.

References

[1] Kofler, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., & Kolomeyer, E. G. Reaction time variability in ADHD: A meta-analytic review of 319 studies. Clinical Psychology Review, 33(6), 795-811, 2013. https://doi.org/10.1016/j.cpr.2013.06.001.

[2] Huang-Pollock, C. L., Karalunas, S. L., Tam, H., & Moore, A. N. Evaluating vigilance deficits in ADHD: A meta-analysis of CPT performance. Journal of Abnormal Psychology, 121(2), 360-371, 2012. https://doi.org/10.1037/a0027205.

[3] Wiley, K., Robinson, R., & Mandryk, R. L. The making and evaluation of digital games used for the assessment of attention: Systematic review. JMIR Serious Games, 9(3), e26449, 2021. https://doi.org/10.2196/26449.

[4] Peñuelas-Calvo, I., et al. Video games for the assessment and treatment of attention-deficit/hyperactivity disorder: A systematic review. European Child & Adolescent Psychiatry, 31, 5-20, 2022. https://doi.org/10.1007/s00787-020-01557-w.

[5] Teruel, M. A., et al. Measuring attention of ADHD patients by means of a computer game featuring biometrical data gathering. Heliyon, 10(5), e26555, 2024. https://doi.org/10.1016/j.heliyon.2024.e26555.

[6] Ruiz-Robledillo, N., et al. Discriminative power of the serious game Attention Slackline in children and adolescents with and without attention-deficit/hyperactivity disorder: Validation study. JMIR Serious Games, 13, e65170, 2025. https://doi.org/10.2196/65170.

[7] Johnson, M. J., & Willsky, A. S. Bayesian nonparametric Hidden Semi-Markov Models. Journal of Machine Learning Research, 14, 673-701, 2013.

[8] Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., & Fernández-Manjón, B. Applications of data science to game learning analytics data: A systematic literature review. Computers & Education, 141, 103612, 2019. https://doi.org/10.1016/j.compedu.2019.103612.

[9] Heller, M. D., et al. A machine learning-based analysis of game data for attention deficit hyperactivity disorder assessment. Games for Health Journal, 2(5), 291-298, 2013. https://doi.org/10.1089/g4h.2013.0058.

[10] Trujillo, J., et al. A multimodal dataset for neurophysiological and AI applications. Scientific Data, 13, 436, 2026. https://doi.org/10.1038/s41597-026-06758-7.

[11] von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. PLOS Medicine, 4(10), e296, 2007. https://doi.org/10.1371/journal.pmed.0040296.

[12] Collins, G. S., et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378, 2024. https://doi.org/10.1136/bmj-2023-078378.

[13] Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. Machine learning algorithm validation with a limited sample size. PLOS ONE, 14(11), e0224365, 2019. https://doi.org/10.1371/journal.pone.0224365.

[14] de Hond, A. A. H., Leeuwenberg, A. M., Hooft, L., et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review. npj Digital Medicine, 5, 2, 2022. https://doi.org/10.1038/s41746-021-00549-7.

[15] Tamm, L., Narad, M. E., Antonini, T. N., O'Brien, K. M., Hawk, L. W. Jr., & Epstein, J. N. Reaction time variability in ADHD: A review. Neurotherapeutics, 9(3), 500-508, 2012. https://doi.org/10.1007/s13311-012-0138-5.

[16] Karalunas, S. L., Geurts, H. M., Konrad, K., Bender, S., & Nigg, J. T. Reaction time variability in ADHD and autism spectrum disorders: Measurement and mechanisms of a proposed trans-diagnostic phenotype. Journal of Child Psychology and Psychiatry, 55(6), 685-710, 2014. https://doi.org/10.1111/jcpp.12217.

[17] Arrondo, G., et al. Systematic review and meta-analysis: Clinical utility of continuous performance tests for the identification of attention-deficit/hyperactivity disorder. Journal of the American Academy of Child & Adolescent Psychiatry, 63(2), 154-171, 2024. https://doi.org/10.1016/j.jaac.2023.03.011.

[18] Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. JMIR Serious Games, 4(2), e11, 2016. https://doi.org/10.2196/games.5888.

[19] Hall, C. L., Valentine, A. Z., Groom, M. J., Walker, G. M., Sayal, K., Daley, D., & Hollis, C. The clinical utility of the continuous performance test and objective measures of activity for diagnosing and monitoring ADHD in children: A systematic review. European Child & Adolescent Psychiatry, 25(7), 677-699, 2016. https://doi.org/10.1007/s00787-015-0798-x.