Trial-level pupil dynamics reflect attention-related variability and medication-associated modulation in ADHD Supplementary Materials

Saumya Yadav, Aadya ., and Jainendra Shukla

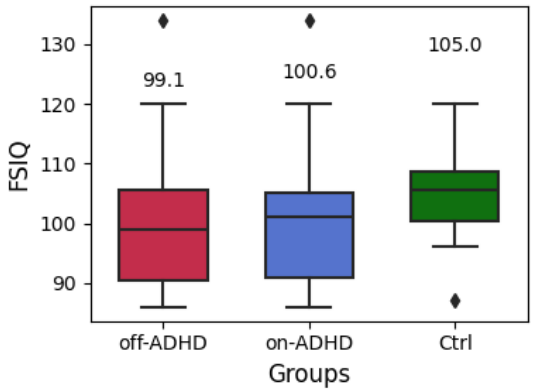

Figure 1: Distribution of Full Scale IQ (FSIQ) scores across diagnostic groups. The control group shows a higher median FSIQ (105.0) than the on-ADHD (100.6) and off-ADHD (99.1) groups. Boxes indicate interquartile range (IQR), with whiskers representing 1.5×IQR and diamonds denoting outliers.

Group	Low IQ (≤89)	Average IQ (90–109)	High IQ (110+)
Ctrl	1	16	5
off-ADHD	6	19	2
on-ADHD	3	12	2

Table 1: Distribution of Participants by Group and FSIQ Category.

Descriptive IQ-stratified results in Table 4. Off-ADHD participants showed numerically higher inter-trial variability than controls within each IQ stratum (Low: 0.466 vs. 0.195; Average: 0.426 vs. 0.386; High: 0.454 vs. 0.374). How-

Hypothesis	Primary comparison or model term	Analysis sample	Raw p	FDR-adjusted q	Significant after FDR
H1	Ctrl vs. Off-ADHD inter-trial variance	Ctrl $n = 22$; Off-ADHD $n = 28$	.037	.051	No
H1	Ctrl vs. Off-ADHD intra-trial variance (SD)	Ctrl $n = 22$; Off-ADHD $n = 28$	.022	.040	Yes[a]
H2	Off-ADHD inter-trial variability–accuracy association	Off-ADHD $n = 28$	.007	.019	Yes
H3a	Off-ADHD vs. On-ADHD mean relative pupil response	$n = 17$ paired participants	.001	.011	Yes
H3b	Off-ADHD vs. On-ADHD inter-trial variance	$n = 17$ paired participants	.031	.049	Yes
H3b	Ctrl vs. On-ADHD inter-trial variance	Ctrl $n = 22$; On-ADHD $n = 17$	.899	.899	No
H4	RT effect in the Off-ADHD reference condition	Trial-level GEE	.022	.040	Yes
H4	On-ADHD × RT interaction	Trial-level GEE	.002	.011	Yes
H4	On-ADHD × RT^2 interaction	Trial-level GEE	.003	.011	Yes
H4	Ctrl × RT interaction	Trial-level GEE	.769	.886	No
H4	Ctrl × RT^2 interaction	Trial-level GEE	.805	.886	No

Table 2: Benjamini–Hochberg false discovery rate correction across the primary inferential tests for H1-H4. Raw p-values are those reported in the main text. FDR significance was evaluated at $q < .05$. *Note: FDR-adjusted values were calculated using the Benjamini-Hochberg procedure across the 11 primary tests.* [a] *Although statistically significant after FDR correction, the intra-trial SD difference was opposite to the direction predicted by H1: Off-ADHD showed lower rather than higher within-trial variability. The H4 RT term is labelled as the effect in the Off-ADHD reference condition, assuming Off-ADHD was the reference group in the GEE model.*

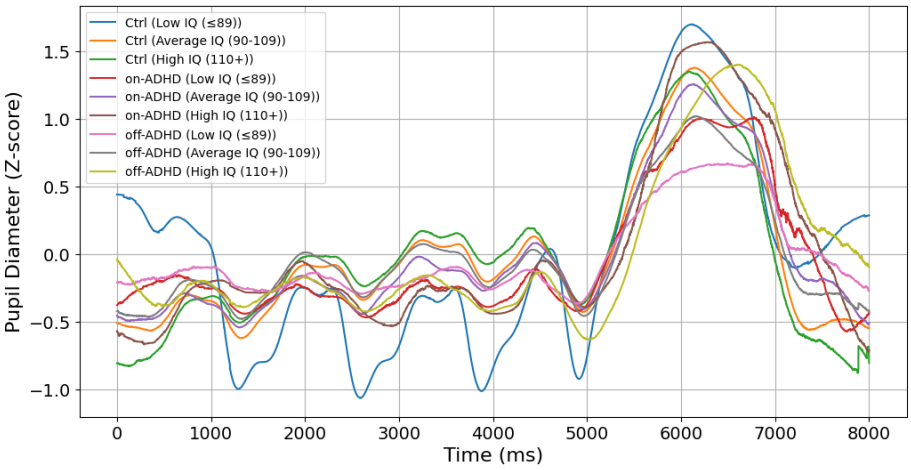

Figure 2: Time-aligned average pupil diameter across diagnostic groups and IQ strata during the working memory task.

ever, the Low-IQ control estimate was based on a single participant and should be interpreted cautiously. Within-trial variability showed less consistent descriptive patterns across IQ strata: Off-ADHD values were lower in the Low-IQ stratum (0.597 vs. 0.833), similar in the Average-IQ stratum (0.732 vs. 0.777), and slightly higher in the High-IQ stratum (0.779 vs. 0.744). Given the small numbers of participants outside the Average-IQ range, these stratified findings are presented as descriptive supplementary analyses rather than confirmatory group comparisons.

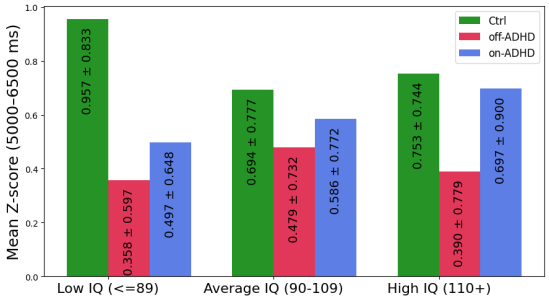

Figure 3: Mean ± SD of pupil diameter during the probe phase (5000–6500 ms) across diagnostic groups and IQ strata.

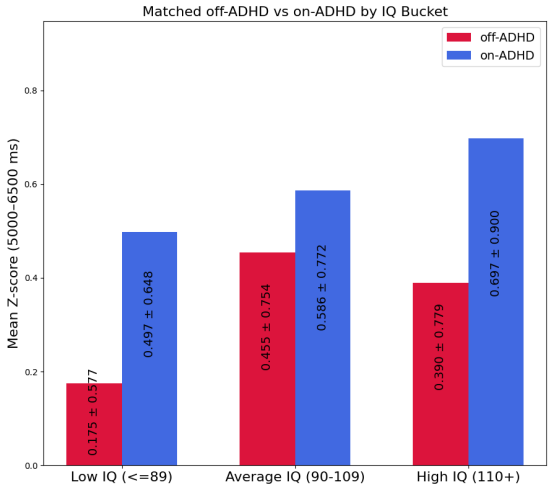

Figure 4: Mean probe-window pupil response for the matched ADHD participants assessed off and on methylphenidate, stratified by IQ group. Bars show group means, and labels indicate mean ± SD. Mean pupil response was higher in the On-ADHD condition across all IQ strata. IQ-stratified values are descriptive because of the small subgroup sizes.

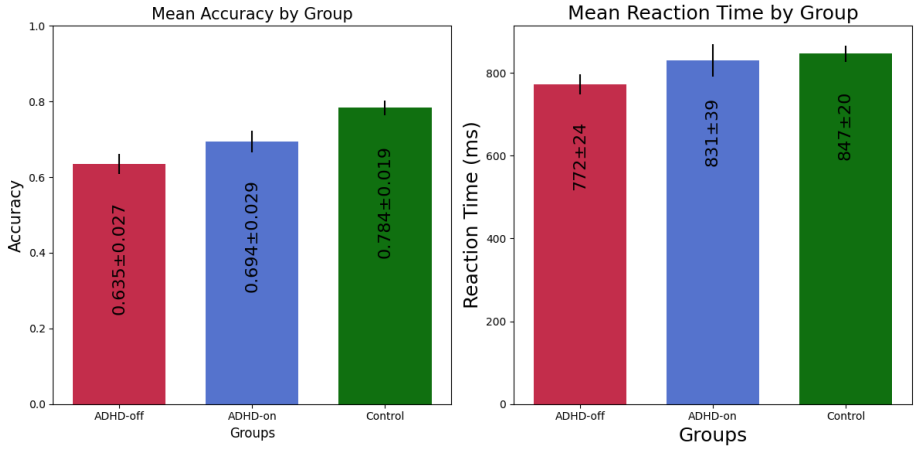

Figure 5: Mean accuracy (left) and reaction time (right) across diagnostic groups. Off-ADHD participants show the lowest accuracy and fastest responses, indicating impulsivity. On-ADHD participants improve in accuracy with slower responses, while controls exhibit the highest accuracy and longest reaction times, reflecting a more balanced speed–accuracy trade-off.

Hypothesis	Metric	Statistical Test	Groups Compared	Table and Figure
H1	Inter-trial variance and within-trial SD	Bayesian hierarchical location–scale models; Mann–Whitney U tests with rank-biserial correlations as participant-level sensitivity analyses	Ctrl vs. Off-ADHD	Table 2 and Figure 2
H2	Inter-trial variance and task accuracy	Spearman rank correlations; permutation comparison of correlations; group-by-inter-trial-variance interaction analysis	Within Off-ADHD and Ctrl, followed by direct comparison between conditions	Figure 3
H3a	Mean probe-window pupil response and pupil-trajectory similarity	Paired-samples t-test; Wilcoxon signed-rank sensitivity analysis; Fréchet distance, dynamic time warping, and bootstrap analysis	Paired Off-ADHD vs. On-ADHD; descriptive comparisons with Ctrl	Tables 2–3 and Figures 2 and 4
H3b	Inter-trial variance	Paired-samples t-test; Wilcoxon signed-rank sensitivity analysis; Mann–Whitney U test for the independent comparison	Paired Off-ADHD vs. On-ADHD; Ctrl vs. On-ADHD	Table 2 and Figure 2
H4	Trial accuracy as a function of reaction time	Logistic regression and GEE with RT, RT^2, condition, and interaction terms; participant-level group comparisons	Ctrl, Off-ADHD, and On-ADHD	Table 4 and Figure 5

Table 3: Summary of the hypotheses, outcome metrics, statistical analyses, and corresponding results. IQ-stratified analyses are reported separately in the Supplementary Material as descriptive and exploratory analyses.

Exploratory gender-stratified analyses

Exploratory gender-stratified analyses indicated that female participants in the Off-ADHD condition ($N = 5$) showed numerically higher inter-trial pupil vari-

Group	IQ Bucket	# Participants	# Trials	Inter-trial variability
Ctrl	Low (≤ 89)	1	152	0.195
Ctrl	Average (90-109)	16	2415	0.386
Ctrl	High (110+)	5	670	0.374
off-ADHD	Low (≤ 89)	6	813	0.466
off-ADHD	Average (90-109)	19	2739	0.426
off-ADHD	High (110+)	2	299	0.454
on-ADHD	Low (≤ 89)	3	456	0.369
on-ADHD	Average (90-109)	12	1794	0.383
on-ADHD	High (110+)	2	302	0.425

Table 4: Inter-trial variability of z-scored pupil diameter during the probe phase (5000–6500 ms), stratified by diagnosis and IQ. Note: 27 participants in off-ADHD

Group	N	Pearson r	p-value	Spearman ρ	Spearman–Brown
Ctrl	22	0.493	0.020	0.458	0.660
off-ADHD	28	0.365	0.057	0.263	0.534
on-ADHD	17	0.412	0.101	0.385	0.583

Table 5: Split-half reliability of inter-trial pupil variability. Reliability was computed by dividing trials into two halves within each participant and correlating variability estimates across halves. Spearman-Brown correction accounts for full-length reliability.

Curve 1	Curve 2	Fréchet Distance
	Top 10 Most Similar	
on-ADHD (Low IQ)	off-ADHD (Average IQ)	0.1459
Ctrl (Average IQ)	on-ADHD (Average IQ)	0.1510
Ctrl (Average IQ)	Ctrl (High IQ)	0.1844
on-ADHD (Average IQ)	off-ADHD (Average IQ)	0.2566
Ctrl (High IQ)	on-ADHD (Average IQ)	0.2899
on-ADHD (Low IQ)	on-ADHD (Average IQ)	0.3080
Ctrl (Average IQ)	on-ADHD (High IQ)	0.3471
on-ADHD (Low IQ)	off-ADHD (Low IQ)	0.3561
Ctrl (Average IQ)	off-ADHD (Average IQ)	0.3679
off-ADHD (Low IQ)	off-ADHD (Average IQ)	0.3823
	Top 10 Most Dissimilar	
on-ADHD (High IQ)	off-ADHD (Average IQ)	0.6566
Ctrl (High IQ)	off-ADHD (High IQ)	0.6612
Ctrl (Low IQ)	off-ADHD (Average IQ)	0.7128
Ctrl (Low IQ)	off-ADHD (High IQ)	0.7282
Ctrl (High IQ)	off-ADHD (Low IQ)	0.7294
Ctrl (Average IQ)	off-ADHD (Low IQ)	0.7437
Ctrl (Low IQ)	on-ADHD (Low IQ)	0.7643
off-ADHD (Low IQ)	off-ADHD (High IQ)	0.8022
on-ADHD (High IQ)	off-ADHD (Low IQ)	0.9149
Ctrl (Low IQ)	off-ADHD (Low IQ)	1.0748

Table 6: Top 10 Most Similar and Most Dissimilar Pupil Curves Based on Fréchet Distance

Comparison	DTW Mean	DTW 95% CI	Fréchet Mean	Fréchet 95% CI
Ctrl vs on-ADHD	2.759	[0.369, 7.566]	0.262	[0.114, 0.439]
Ctrl vs off-ADHD	9.233	[4.240, 14.181]	0.480	[0.289, 0.664]
on-ADHD vs off-ADHD	6.115	[1.053, 12.078]	0.353	[0.138, 0.579]

Table 7: Bootstrap distances between group-mean pupil trajectories. Within each of 1,000 bootstrap iterations, trajectories were resampled within each group, group-mean trajectories were calculated, and DTW and Fréchet distances were computed between the resulting means. These values are therefore different from the pairwise trajectory distances reported in the main manuscript.

Comparison	Observed value	Bootstrap mean	95% CI
Ctrl-Off-ADHD DTW	9.2636	9.1569	[4.2165, 14.0462]
Ctrl-On-ADHD DTW	2.0702	2.7553	[0.3299, 7.5347]
ΔDTW	7.1934	6.4016	[0.3030, 11.7093]

Table 8: Direct bootstrap contrast of DTW distances between group-mean trajectories. (Note: Two-sided bootstrap $p = .038$ for $\Delta\text{DTW} = D(\text{Ctrl}, \text{Off}) - D(\text{Ctrl}, \text{On})$

.

Group	**IQ Bucket**	**Participants**	**Trials**	**Inter-trial variability**
off-ADHD	Low ($\leq$ 89)	3	429	0.506
off-ADHD	Average (90-109)	12	1751	0.423
off-ADHD	High (110+)	2	299	0.454
on-ADHD	Low ($\leq$ 89)	3	456	0.369
on-ADHD	Average (90-109)	12	1794	0.383
on-ADHD	High (110+)	2	302	0.425

Table 9: Within-subject comparison of inter-trial pupil variability between off-medication and on-medication conditions in the crossover subset (N = 17). Each row represents a single participant.

ance than male participants ($N = 22$; 0.479 ± 0.113 vs. 0.341 ± 0.101). In the On-ADHD condition, the difference was smaller, with values of 0.316 ± 0.076 for females ($N = 3$) and 0.295 ± 0.097 for males ($N = 14$). By contrast, female Ctrl participants ($N = 4$) showed slightly lower inter-trial variance than male Ctrl participants ($N = 18$; 0.279 ± 0.088 vs. 0.294 ± 0.085; Figure 6). These descriptive patterns are broadly consistent with prior evidence suggesting gender-related differences in ADHD presentation and arousal regulation [1–4]. However, the small and highly unbalanced subgroup sizes, particularly the three female participants in the On-ADHD condition, preclude reliable inferential interpretation. These findings should therefore be considered hypothesis-generating and require confirmation in adequately powered and demographically balanced samples.

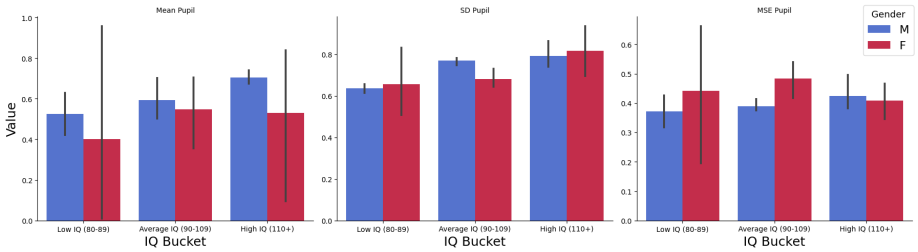

Figure 6: Gender-wise Mean, Intra-trial variability (SD), and Inter-trial variability of Pupil Responses across IQ Buckets (5000–6500 ms window) in the on-ADHD group.

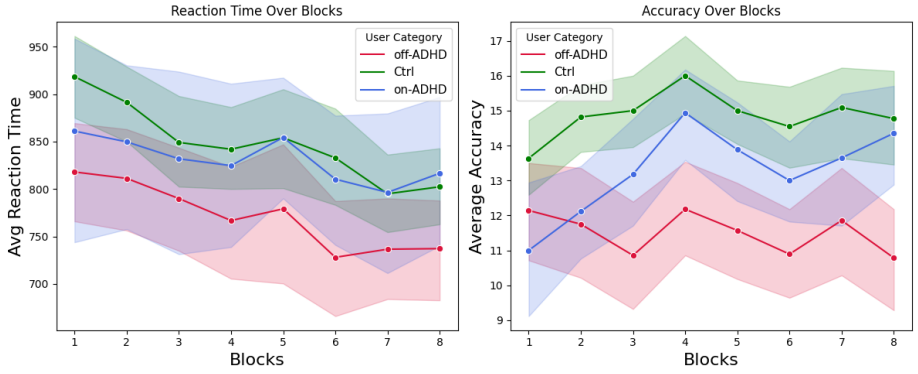

Figure 7: Block-wise trends in RT (left) and accuracy (right) across the eight task blocks for each diagnostic group.

References

[1] Jessica W O'Brien, Lauren R Dowell, Stewart H Mostofsky, Martha B Denckla, and E Mark Mahone. Neuropsychological profile of executive function in girls with attention-deficit/hyperactivity disorder. *Archives of Clinical Neuropsychology*, 25(7):656–670, 2010.

[2] Patricia O Quinn and Manisha Madhoo. A review of attention-deficit/hyperactivity disorder in women and girls: uncovering this hidden diagnosis. *Prim Care Companion CNS Disord*, 16(3):PCC–13r01596, 2014.

[3] M De Jong, DSMR Wynchank, E Van Andel, ATF Beekman, and JJS Kooij. Female-specific pharmacotherapy in adhd: premenstrual adjustment of psychostimulant dosage. *Frontiers in psychiatry*, 14:1306194, 2023.

[4] Frédéric Dehais, Alex Lafont, Raphaëlle Roy, and Stephen Fairclough. A neuroergonomics approach to mental workload, engagement and human performance. *Frontiers in neuroscience*, 14:268, 2020.