Investigating the Effects of Feedback Modality on Compensatory Tracking Performance for Remote Surgery Applications

Jessica Fu[†], Sara Parvaresh Rizi[‡], and Aarya Shah[¶]

[†] jessica.fu1@mail.utoronto.ca | 1011048607

[‡] s.parvareshrizi@mail.utoronto.ca | 1010913451

[¶] aarya.shah@mail.utoronto.ca | 1010871157

Abstract

In remote surgery applications, the loss of direct haptic feedback necessitates sensory substitution to convey critical information such as instrument depth [1]. This study compares symbolic "numerical" (Level A) and sensory "spatial-colour" (Level B) visual feedback on compensatory tracking performance within the context of Robot-Assisted Minimally Invasive Surgery (RAMIS). Using a sample of 32 participants, performance was quantified through completion runtime and accuracy (pixel area deviation).

While initial analyses ($N = 32$) showed that participants were 8.55s faster on average with spatial-colour feedback, this result was statistically significant under the directional one-tailed test ($p = 0.0293$ for time). Accuracy was not significantly better for numerical feedback under its directional one-tailed test ($p = 0.6005$), and the observed mean difference was opposite to **Hypothesis 2.2** (numerical had 27.9 px^2 larger area off target on average). A sensitivity analysis ($N = 28$) omitting four technical outliers similarly showed a statistically significant speed advantage of 8.14s for spatial-colour feedback ($p = 0.0165$), while the accuracy-direction hypothesis remained unsupported ($p = 0.9054$). Qualitative survey data indicated that the high cognitive load required to decode symbolic numerical data was perceived as "distracting" and "mentally taxing," negating its theoretical precision benefit. Furthermore, a strong negative correlation ($r \approx -0.75$) confirmed a robust speed-accuracy trade-off across both modalities. Analysis of gender and video-gaming covariates showed no significant differences in mean performance, though video gamers demonstrated a more pronounced speed-accuracy trade-off and reported higher utility for spatial-colour cues. These findings suggest that surgical robotic interfaces should prioritize pre-attentive sensory-spatial cues to minimize operator lag time while maximizing accuracy.

Github Repository: `https://github.com/snowxzf/MIE_Project`

1 Introduction

Cognitive ergonomics and workload relate to the ability to process sensory input, such as sight, sound, and touch, into motor responses and 'working memory' to complete tasks [2]. Surgeons performing high-precision tasks like robotic surgeries use cognitive training with "cognition to integration to autonomous" learning stages. Here, the brain consistently applies conscious understanding into deliberate motor execution until using little monitoring as the surgeon reaches an 'autonomous' stage [3].

In Robot-Assisted Minimally Invasive Surgery (RAMIS), the physical separation between the surgeon and the patient disrupts direct haptic (touch) feedback [4]. This absence of tactile information necessitates sensory substitution, where critical data like instrument depth must be translated into visual cues. While cognitive ergonomics research has explored how workload affects performance, there is a lack of comparative data on which visual modalities most effectively bridge this 'depth-sensing' gap [5, 6] by optimizing accuracy while minimizing lag time, making our research both novel and significant.

This begets the question: *How do Symbolic (Level A) and Sensory (Level B) feedback modalities compare in accuracy and runtime when performing a continuous tracking task, simulating surgery?*

Symbolic feedback uses characters like Arabic numerals, relying on foveal vision and "semantic interpretation," thus requiring more working memory and focal attention. This often causes "in-attentional blindness" [7, 8]. Meanwhile, sensory feedback uses spatial-colour gradients, which target the peripheral vision, excelling at pre-attentive motion and colour detecting [9]. Visual metrics were prioritized for their efficacy in surgeries, as acoustic factors are more distracting [10], and 90% of feedback to the brain is visual [11], making it more accessible. Spatial-colour and numerical feedback are distinct, yet immensely popular as they compare qualitative and quantitative feedback [12, 13]. Numbers best and most accurately quantify trends and compare to a baseline, whereas spatial-colour uses selective attention, processing 60 000 times faster than text and reducing cognitive load compared to reading numbers [11].

This report highlights the methods and biases, alongside course concepts used to analyze and infer the influence of covariants on data. The result implications were applied to a surgical application.

2 Methodology

The independent variables are the two feedback modalities, namely Level A and Level B feedback. The dependent variables that will be tested are runtime (time taken to complete the task) and accuracy (how close the individuals were to the line). Accuracy will be measured by calculating the area bounded between the drawn curve and the target curve. This section highlights our key design decisions, biases and limitations in performing this study.

2.1 Participants and Sampling

A sample size of 32 University of Toronto (UofT) second-year Engineering Science (EngSci) students were taken to perform this experiment: 16 Females and 16 Males between ages 19 to 21. To mitigate

potential order effects and task habituation, a counterbalanced experimental design was employed [14]. Within each gender cohort ($N = 16$), participants were equally divided into two groups: eight individuals performed the numerical modality (Level A) first, while the other eight began with the spatial-colour feedback (Level B). This distribution ensures that any observed improvements in runtime or accuracy are attributable to the feedback modality itself, rather than a learning effect from tracing the same pattern [15].

The study was divided by gender and treated as a covariant because studies claim that fine motor task performance depends on age and test conditions, with Males between the age of 12 and 17 outperforming Females in fine motor tasks, although there is a weak significance to confirm this [16].

The $N = 32$ sample was taken to satisfy both the theoretical and structural requirements of the experimental design. A sample size exceeding the $N > 30$ threshold was necessary to use Central Limit Theorem (CLT) and assume normality across the entire population, thus strengthening the validity of the parametric inferential statistical analysis [17], while allowing for an equitable distribution of 8 Males and Females performing Level A or B feedback first.

A participation form was given to all individuals to assess any covariants and diagnose that can hinder their performance (ex. fine motor issues, colourblindness). This survey is available in Appendix A.

While UofT EngSci's were chosen for accessibility, they cannot replicate the experience of a surgeon, prompting selection bias. Thus, playing video games became a covariant as it best improves cognitive training and hand-eye coordination, simulating the skills of surgeons [18], while being an accessible fine motor task for most engineering students. Time covariates are ignored as accuracy was prioritized. However, a limitation is the uneven split in the amount of time participants spent gaming, where 17 participants spent 0 hours/day, while 15 participants spent between 1-4 hours/day (>0 hours/day).

The survey used objective questions to limit speculation. However, factors like the number of hours of gaming can fluctuate throughout a semester.

2.2 Experimental Apparatus

Appendix B provides a script that was used to convey the same information to each participant. Each test was conducted in a quiet area by participants during their booked time slot. They were asked if they preferred Mac or Windows due to the dexterity one may face when operating a different type of laptop. Both devices were 12 inches long [19,20], had a trackpad sensitivity and cursor speed of 5, maximum brightness and had the test maximized. Users were prohibited from zooming in or using a touchscreen. Figure 1 is the tested curve for all participants for both modalities.

Figure 1: The tested curve presented on the UI.

The curve in Figure 1, and found on Github combines common incision geometries, like midline, coronal, and transverse incisions [21] to simulate different surgeries. As re-incisions are prevalent, where in 66% surgeries, doctors reoperate on the initial incision to minimize the amount of scarred skin, they must be proficient with repetitive pattern tracing [22]. This encouraged our curve to have repetitive patterns. The same curve was tested for both feedback modes to maintain the 'complexity' of the curve.

All participant data were downloaded as a .csv file and compiled into `trial\_metrics\_summary.csv` and `participant\_order.csv`. After each trial, participants were asked questions, found in Appendix A, about their opinions on both modalities and if they experienced a learning curve from subsequent trials.

2.3 Experimental Task & Procedure

Participants were given four practice trials of randomly generated curves without feedback [23] to become familiar with test performance, difficulty and navigation. However, they were not required to use all their practice trials if they felt ready. To increase comfort, the user could trace the curve without having to constantly press down on the trackpad.

All users were told they were being timed with a stopwatch for both assessments, but were hiding it to encourage them to focus on accuracy.

Appendix C shows how the UI provides feedback for each modality. Instructions for how the UI communicated feedback were both written on the top of the test box and verbally communicated from the script. In the numerical test, a decimal number following the cursor would alert the user on how many pixels they deviated. The goal was to stay as close to 0.0 for accuracy. Meanwhile, the spatial-gradient modality had a constant light green background to indicate "closeness" to the curve. The gradient changed every 2.5 pixels to ensure it was sensitive enough for small deviations while not being distracting, as is the case for 1 to 2 pixel deviations [24]. It progresses from green to red the farther one strays from the line.

3 Analyses

This section plans to use t-distribution, T-tests, Pearson correlation, sensitivity analysis, normality and inferential statistics to test our hypotheses.

3.1 Hypotheses

This study aims to test the following 4 hypotheses.

Hypothesis 1 (speed-accuracy trade-off): There will be a negative correlation between runtime and accuracy due to the Speed-Accuracy tradeoff, for both feedback modalities. As the participant attempts to finish the test faster (speed), they will likely move hastily, sacrificing their accuracy. Note that a highly positive correlation indicates that increased runtime would lead to a higher area between the drawn and generated curves, and thus that the participants had lower accuracy the more time they spent. Consequently, a negative correlation is preferred. This is written more in depth in Section 3.4.

Using the Cognitive Load Theory [25], numerical values require more mental resources to understand, but they also provide a more specific and quantifiable metric compared to the spatial gradient. Therefore,

the following two hypotheses were created:

- **Hypothesis 2.1**: Spatial feedback will take less time to finish than numerical feedback.
- **Hypothesis 2.2**: Numerical feedback will yield higher accuracy than spatial-colour feedback.

Hypothesis 3 (male vs. female performance): Using the evidence in Section 2.1, male participants will outperform female participants overall. This hypothesis assesses the gender covariant in the study.

Video games provide sufficient cognitive training to improve hand-eye coordination [26] through colour-feedback, allowing them to prioritize accuracy regardless of time. This thus makes it a covariate for this study. There are two hypotheses on the results of gamers vs. non-gamers below:

- **Hypothesis 4.1**: Participants who spend more time playing video games will have a high accuracy independent of any time-accuracy tradeoff, especially with spatial feedback.
- **Hypothesis 4.2**: Participants who play more video games will have shorter response times in both modalities, as they are more used to completing fine-motor tasks.

For **Hypothesis 3 and 4.1/4.2**, to assess the magnitude and direction of potential variables that impact cognitive training, we found the correlation between the runtime-accuracy ratio for Females and compared it to that of Males for both feedback modalities. Similarly, we assessed the correlation between individuals who played video games and those who did not.

The null hypothesis for testing these assumptions is that there is no correlation between any independent and dependent variables.

3.2 Normality Assumption Checks

All data was analyzed using R, and is found in Appendix F. Normality tests were conducted using:

- Shapiro-Wilk (W)for a strict statistical test of Gaussian distribution [27] (Appendix F.1).
- Quantile-quantile plots for a qualitative visualization of tail deviations [28] (Appendix F.2).

A QQ plot visually depicted the graphical normality in the different categories, whereas W quantitatively reinforced the normality using analytical approaches. The W test is the most reliable because it is used for sample sizes $N < 50$ [29, 30].

3.3 Use of T Distributions

A one-tailed paired test was used for the primary modality hypotheses because **Hypothesis 2.1** and **Hypothesis 2.2** are directional. Specifically, for completion time we tested whether numerical minus spatial-colour was greater than zero (*spatial-colour faster*), and for area off target we tested whether numerical minus spatial-colour was less than zero (*numerical more accurate*).

A t-distribution is valid for small sample sizes, ideally $N < 30$, although it can be used for $N \geq 30$ because we had unknown population standard deviations [31]. Thus, the t-distribution was used to extrapolate the population mean and confidence interval [32]. Using the standard $\alpha = 0.05$ significance

interval, if the one-tailed p-value was equal to or less than α, then the null hypothesis was rejected in the prespecified direction. The code for the directional one-tailed test can be found in Appendix F.3.

Calculations for mean and standard deviation code can be found in Appendix F.7. For comparisons of mean completion time (s) and mean area off target between independent groups—specifically, Females versus Males(**Hypothesis 3**), and participants below versus above the sample median of self-reported gaming hours per day (**Hypothesis 4.1/4.2**)—we used Welch's two-sample t-test rather than Student's classical two-sample t-test. The code for Welch's t-test can be found in Appendix F.4. The Welch procedure tests the same null hypothesis (equality of population means) but does not assume equal variances between groups; it uses the separate sample variances and adjusts the degrees of freedom (often non-integer) to account for heteroscedasticity [33]. This choice is standard when group sizes or spread may differ, which is plausible for self-report and performance subgroups.

3.4 Assessing the Speed-Accuracy Tradeoff

The code for calculating the Pearson's coefficient can be found in Appendix F.5. We use Pearson's r to summarize the linear association between runtime and accuracy (and between Δ time and Δ area) on the same participants. It ranges from -1 to $+1$ and is scale-free [34]. **Hypothesis 1** is framed as a linear speed–accuracy relationship, so Pearson is appropriate. Other quantities like Spearman rank correlation would stress monotonicity rather than linearity [35]. A negative r between time and area off target indicates that slower trials tend to be more accurate. Inference assumes roughly bivariate normality; where subgroup normality is borderline, we treat r as descriptive support alongside scatter plots and QQ checks, not as a strict distributional proof.

3.5 Conducting a Sensitivity Analysis

A sensitivity analysis allows us to understand how different independent variables and data points influence the dependent variable and overall data analysis, making them useful for determining outliers and deepening our understanding of the influence of independent variables by making assumptions [36].

Data cleaning followed a protocol where outliers were defined using a conservative criterion of $|Z - score| > 3$ for standard deviations or values exceeding 1.5 × IQR. This threshold ensured that the sensitivity analysis excludes only extreme technical anomalies, thereby preserving the integrity of the core dataset for hypothesis testing [37]. We compared performance with and without outliers to verify the robustness of our findings. The above analysis (z-score and IQR) was implemented in our code and can be seen in Appendix F.6.

4 Results

This section outlines assumptions in determining normality, if the data aligns with the hypotheses, and an assessment of covariances. Table 1 gives a comprehensive summary of the primary inferential statistics, including mean, standard deviation, paired t-test outcomes and Pearson correlation coefficients for both the full ($N = 32$) and filtered ($N = 28$) datasets. All

Table 1: Full ($N = 32$) vs. outlier-excluded ($N = 28$). Descriptive rows are mean (SD) across participants in each feedback mode. Paired differences are numerical minus spatial-colour (Level A − Level B): a positive time difference means slower completion under numerical feedback; a positive area difference means larger area off target under numerical feedback (worse accuracy).

Metric	**Full ($N = 32$)**	**Filtered ($N = 28$)**
Descriptive means (SD)		
Time, numerical (s)	85.50 (38.80)	82.80 (34.20)
Area, numerical (px^2)	3060 (874)	3100 (871)
Time, spatial-colour (s)	76.9 (30.6)	74.6 (27.6)
Area, spatial-colour (px^2)	3030 (882)	2980 (675)
Paired t-tests on mean participant difference (A − B)		
Time mean diff. (s)	8.55 ($p = 0.029$)	**8.14 ($\mathbf{p = 0.017}$*)**
Area mean diff. (px^2)	27.90 ($p = 0.601$)	124 ($p = 0.905$)
Pearson correlation (r)		
Num. (time vs. area)	$r = -0.747$ ($p < 0.001$)	$r = -0.785$ ($p < 0.001$)
Spatial (time vs. area)	$r = -0.718$ ($p < 0.001$)	$r = -0.734$ ($p < 0.001$)
ΔTime vs. ΔArea	$r = -0.520$ ($p = 0.002$)	$r = -0.694$ ($p < 0.001$)

**One-tailed paired t-test (direction: numerical > spatial-colour for time) significant at $\alpha = 0.05$.*

data is reported to three significant figures [38]

4.1 Normality Assumption Checks

Figure 9 in Appendix D shows a quantile-quantile (QQ) plot of the data for Level A and B feedback for each variable (time and accuracy). Qualitatively, the data follows the linear trend of the red line, with minor standard deviations, showing normality. The R code for the QQ plots is detailed in Appendix F.2.

To quantify this normality, Table 6 in Appendix D outlines the Shapiro-Wilk statistics (W). As the W values are close to 1, they point towards normality in the data [39]. Numerical runtime and spatial-colour accuracy showed formal non-normality via the Shapiro-Wilk test ($p < 0.05$). However, because this test can be overly sensitive to minor outliers in moderate samples, a triangulated approach was used [37]. Visual inspection of the QQ plots revealed a major linear trend, suggesting the data are sufficiently normal for parametric analysis. Consequently, we maintained the normality assumption for the t-tests, though we acknowledge these specific metrics represent a potential limitation in the inferential model.

4.2 Feedback Modality Testing

Figure 2 provides a box plot analyzing how completion time changed by feedback type. As seen quantitatively in Table 1, participants were slower with numerical feedback than spatial-colour by about 8.55 s on average, aligning with **Hypothesis 2.1**. Under the directional one-tailed paired test for time ($\mu_{\text{numerical−spatial}} > 0$), the p-value was 0.029, below $\alpha = 0.05$, so **Hypothesis 2.1** was supported.

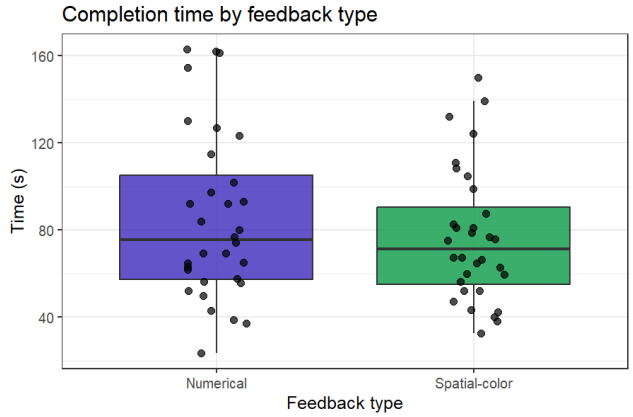

Figure 2: Box Plot for Completion time based on numerical (purple) and spatial-colour (green) feedback. Qualitatively, there is less variance in time among participants in spatial-gradiant than in numerical.

Figure 3 shows a box plot between numerical and spatial colour for accuracy, measured by area off of the target curve.

From Table 1, the area off target was larger in numerical than spatial-colour by 27.9 px^2, opposite to **Hypothesis 2.2**. Under the directional one-tailed paired test for accuracy ($\mu_{\text{numerical}-\text{spatial}} < 0$), the p-value was 0.601, which is greater than $\alpha = 0.05$; therefore, **Hypothesis 2.2** was rejected.

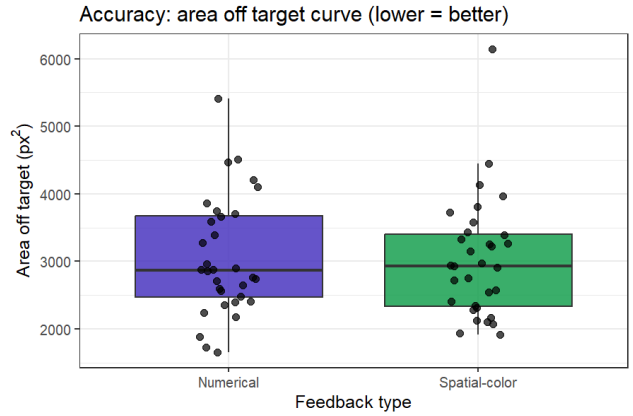

Figure 3: Box Plot for performance accuracy, measured by the amount of deviation the individual had from the target curve. The lower the score, the better. Both modes had similar variability and accuracy scores, even though spatial-colour provided better accuracy, disproving **Hypothesis 2.2**.

The qualitative survey results (Figure 23 in Appendix G) provide a cognitive basis for the observed quantitative trends. While numerical feedback was intended to provide high-precision metrics, 50% of participants ($N = 16$) categorized it as "Distracting" or "Stressful," often noting that the rapid decimal changes forced a transition from intuitive tracing to deliberate focal attention. This increased cognitive load likely accounts for the higher variance and slower runtimes observed in the numerical trials. In contrast, 59% of participants ($N = 19$) explicitly preferred the spatial-colour modality, describing the light-green gradient as "calming" and "intuitive." This aligns with the theory of pre-attentive processing; by allowing users to monitor error via peripheral vision, the spatial-colour mode minimized "inattentional blindness" and facilitated a smoother motor "flow." Additionally, the 69% perceived learning rate ($N = 22$) confirms that despite counterbalancing, the majority of participants successfully internalized the sinusoidal pattern, highlighting the role of habituation in repetitive surgical tasks.

4.3 Correlation for Speed-Accuracy Trade-off

The correlation between runtime and accuracy for both feedback modalities were negative with low p-values as seen in Table 1. The magnitude of these coefficients—ranging from $r = -0.72$ to -0.80 across all subgroups—indicates a strong linear relationship between speed and accuracy. This confirms **Hypothesis 1**, where participants, regardless of modality, consistently modulated their pace to maintain precision, a hallmark of the Speed-Accuracy Trade-off.

4.4 Outlier Testing & Sensitivity Analysis

Upon screening the data for outliers, as mentioned in Section 3.5, we identified four participants as outliers due to extreme deviations in completion time or area error that likely reflected technical hesitation rather than modality effects.

To confirm our findings, a sensitivity analysis was conducted by comparing results from the full dataset ($N = 32$) against the filtered dataset ($N = 28$) [40]. The normality of the outliers was assessed in Table 7 in Appendix D, which shows that while the score for normality improved for numerical accuracy compared to the full sample, there is still a $p < 0.05$ for spatial-colour runtime. However, the QQ plots (Figure 10 in Appendix D), imply normality, they will be cautiously used for smaller sample sizes. This p-value may be due to a natural limit in human motor precision, where individuals cannot significantly improve accuracy beyond a certain threshold regardless of feedback type [41].

As seen in Table 1, omitting these outliers shifted the one-tailed paired test p-value for completion time from 0.029 to 0.017, strengthening the evidence for a speed advantage with spatial-colour feedback. For accuracy, the one-tailed p-value increased from 0.601 to 0.905, remaining non-significant and opposite to the hypothesized direction. This suggests that the speed advantage of sensory-spatial feedback is robust once extreme technical variances are controlled, whereas the directional numerical-accuracy hypothesis is not supported in this sample.

The correlation analysis remains consistent across both the full and filtered datasets, confirming a strong inverse relationship between runtime and area error as stated in **Hypothesis 1**. Notably, the correlation between the difference in time and the difference in area error strengthened from $r = -0.520$ to $r = -0.694$ ($p < 0.001$) after outlier removal, reenforcing a consistent trade-off.

4.5 Exploratory Demographic Analysis

4.5.1 Impact of Gender on Performance

Figure 12 in Appendix E provides a box plot that compares the performance of Males and Females in terms of speed and accuracy. QQ plots for this sample (Figure 11 in Appendix E) show mostly normal distribution, with full W values (in Table 8, Appendix E), spatial-color for Females should be taken cautiously as it showed a significant departure from normality. Descriptive statistics are summarized in Table 2. On average, both groups were faster with spatial-colour than with numerical feedback. Mean times and areas were similar across genders within each mode.

To test whether mean performance differed between genders (independent groups), we used Welch two-sample t-tests (Table 9, Appendix E). None of the comparisons reached statistical significance at $\alpha = .05$, so we did not find evidence that Males and Females differed in mean time or mean accuracy for either feedback type, nor in the participant-level paired differences (numerical $-$ spatial) for time or area, therefore we reject **Hypothesis 3**.

Table 2: Mean completion time (s) and mean area off target (px^2) by gender ($N = 16$ both) and feedback type. Results are formatted as: *mean* (*standard deviation*)

Gender	**Level**	**Time** (s)	**Area off target** (px^2)
Females	A	86.9 (38.1)	3010 (731)
	B	79.2 (33.2)	3040 (981)
Males	A	84.1 (40.7)	3100 (1020)
	B	74.6 (28.7)	3010 (803)

As well, there was a strong negative Pearson correlation between completion time and accuracy for both genders and modalities. Table 3 showcases the Pearson correlation for Females and Males in both feedback types. This consistent negative relationship indicates that both Males and Females exhibited a similar speed-accuracy trade-off, as shown in the analysis above, thereby matching **Hypothesis 1**.

Overall, these results suggest no correlation between gender and performance, as both groups show similar trends in time and accuracy.

Table 3: Pearson correlation coefficients (r) with their corresponding p value (all within $\alpha = 0.05$ by feedback type (Level A and B) and group (male and female)

Group	Level	r	p
Females	A	−0.80	$< .001$
	B	−0.67	.004
Males	A	−0.72	.002
	B	−0.79	$< .001$

4.5.2 Impact of Gaming Experience

Figure 14 in Appendix E compares participants below vs. above the sample median of self-reported gaming hours per day (approximately 0 vs. $>$ 0 h/day). QQ plots for this sample (Figure 13 in Appendix E) show mostly normal distribution, with full W values (in Table 10, Appendix E) though spatial-color accuracy for 0h/day and spatial-color runtime for >0h/day should be taken cautiously. Descriptive statistics are in Table 4.

Table 4: Mean completion time (s) and mean area off target (px^2) by gaming group and feedback type. $N = 17$ for ∼0 h/day; $N = 15$ for > 0 h/day per mode. Results are formatted as: *mean* (*standard deviation*)

Group	Level	Time (s)	Area off target (px^2)
∼0 h/day	A	86.9 (35.0)	2840 (732)
	B	83.4 (31.3)	2880 (1010)
>0 h/day	A	85.5 (45.1)	3300 (1010)
	B	70.3 (30.0)	3220 (730)

Both groups were faster on average with spatial-colour than numerical feedback. This may be due to the heavy usage of color in several games to give feedback to the gamer in their in-game performance, matching **Hypothesis 4.1** as participants who gamed completed the test much faster with spatial feedback compared to numerical feedback.

However, Welch t-tests (Table 11, Appendix E) did not show statistically significant differences between gaming groups in mean time or mean area for either feedback type, nor in the participant-level paired differences (all $p > .15$). Thus we cannot claim a reliable between-group advantage tied to gaming hours

on these tests, and reject **Hypothesis 4.1**.

Pearson correlations between time and area remained strongly negative within each gaming group and feedback type (Table 5), again aligning with **Hypothesis 1** (a consistent negative linear association).

Table 5: Pearson correlation coefficients (r) with their corresponding p value (all within $\alpha = 0.05$ by feedback type (Level A and B) and group (gaming time)

Group	Level	r	p
∼0 h/day	A	−0.737	.0007
	B	−0.706	.002
>0 h/day	A	−0.790	.0008
	B	−0.739	.002

5 Conclusion

This study compared symbolic numerical and sensory-spatial visual feedback to identify the optimal modality for high-precision compensatory tracking in remote surgery. Our results confirm a fundamental **speed-accuracy trade-off (Hypothesis 1)**, where participants across all demographics modulated their pace to maintain precision ($r \approx -0.750$).

While the theoretical precision of numerical feedback (Level A) was hypothesized to improve accuracy, this was rejected (**Hypothesis 2.2**, one-tailed $p = 0.601$ in the full sample; $p = 0.905$ after outlier exclusion); instead, participants reported significant cognitive overload when decoding symbolic data. Conversely, sensory-spatial feedback (Level B) demonstrated a robust speed advantage in both the full sample and sensitivity analysis (**Hypothesis 2.1**, one-tailed $p = 0.029$ for $N = 32$; $p = 0.017$ for $N = 28$). Furthermore, exploratory analyses rejected gender and gaming experience as significant performance determinants (**Hypotheses 3, 4.1, and 4.2**), suggesting that the advantages of spatial cues are largely universal and independent of prior fine-motor training.

These findings have direct implications for the design of RAMIS interfaces. For real-time tasks like catheter navigation or laser ablation, sensory-spatial cues should be prioritized as they are processed **pre-attentively**. This allows surgeons to utilize peripheral vision for depth monitoring without diverting focal attention from the surgical site, reducing corrective lag time and enhancing patient safety [1].

Future research should transition from the "pattern memorization" observed with sinusoidal paths to more ecologically valid, stochastic trajectories that replicate complex surgical incisions. A longitudinal, multimodal approach—integrating spatial-colour cues with redundant haptic feedback—could further delineate the transition from the cognitive to the autonomous learning stage. By controlling for hardware-specific variables like trackpad resolution and operator posture, subsequent iterations will provide a more granular optimization of the human-machine interface in robotic surgery.

References

[1] B. T. Bethea, A. M. Okamura, M. Kitagawa, and D. D. Yuh, "Application of Haptic Feedback to Robotic Surgery," *Journal of Laparoendoscopic & Advanced Surgical Techniques*, vol. 14, pp. 191–195, Jun 2004.

[2] S. W. Wong and P. Crowe, "Cognitive ergonomics and robotic surgery," *Journal of Robotic Surgery*, vol. 18, 03 2024.

[3] "Understanding motor learning stages improves skill instruction."

[4] M. Friebe, "Palpation sensing for robotic-assisted surgery," *Die Orthopädie*, vol. 55, pp. 3–10, 12 2025.

[5] M. Kitagawa, D. Dokko, A. M. Okamura, and D. D. Yuh, "Effect of sensory substitution on suture-manipulation forces for robotic surgical systems," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 129, pp. 151–158, 01 2005.

[6] D. Liu, K. Zang, and J. Shen, "A shallow–deep feature fusion method for pedestrian detection," *Applied Sciences*, vol. 11, p. 9202, 10 2021.

[7] Y. Cai, S. Hofstetter, B. M. Harvey, and S. O. Dumoulin, "Attention drives human numerosity-selective responses," *Cell Reports*, vol. 39, p. 111005, 06 2022.

[8] A. J. Estudillo, "Exploring the role of foveal and extrafoveal processing in emotion recognition: A gaze-contingent study," *Behavioral Sciences*, vol. 15, pp. 135–135, 01 2025.

[9] I. D. Foundation, "Preattentive visual properties and how to use them in information visualization," 2019.

[10] H. M. Mentis, A. Chellali, K. Manser, C. G. L. Cao, and S. D. Schwaitzberg, "A systematic review of the effect of distraction on surgeon performance: Directions for operating room policy and surgical training," *Surgical endoscopy*, vol. 30, p. 1713–1724, -5 2016.

[11] K. Tsitsirikos, "Beyond the numbers: The hidden power of visual communication in data - datwin's blog," March 12 2025.

[12] T. Bryden, "Qualitatives vs. quantitatives feedback – speak ai," -11-25 2022.

[13] B. Trinh, "Balance subjective and objective feedback for better performance reviews."

[14] D. Albudaiwi, "The SAGE encyclopedia of communication research methods," vol. 4, SAGE Publications, Inc., 2018.

[15] P. C. Price, R. S. Jhangiani, and I.-C. A. Chiang, "Experimental Design," in *Research Methods in Psychology*, BCcampus, 2nd canadian ed., Oct 2015. BC Open Textbook Project.

[16] P. Moreno-Briseño, R. Díaz, A. Campos-Romo, and J. Fernandez-Ruiz, "Sex-related differences in motor learning and performance," *Behavioral and Brain Functions : BBF*, vol. 6, p. 74, -12-23 2010.

[17] A. Ganti, "What Is the Central Limit Theorem (CLT)?." `https://www.investopedia.com/terms/c/central_limit_theorem.asp`, Oct 2024. Investopedia.

[18] F. Yuan, E. Klavon, Z. Liu, R. P. Lopez, and X. Zhao, "A systematic review of robotic rehabilitation for cognitive training," *Frontiers in Robotics and AI*, vol. 8, p. 605715, -5-11 2021.

[19] Apple, "Identify your macbook air model."

[20] "Product specifications - yoga 7/7i 2-in-1 series,"

[21] L. A. Jelinek, M. Marietta, and M. W. Jones, *Surgical Access Incisions*. Treasure Island (FL): StatPearls Publishing, 2026.

[22] C. Strik, M. W. J. Stommel, L. J. Schipper, H. van Goor, and R. P. G. ten Broek, "Risk factors for future repeat abdominal surgery," *Langenbeck's Archives of Surgery*, vol. 401, no. 6, p. 829–837, 2016.

[23] C. R. James, J. A. Herman, J. S. Dufek, and B. T. Bates, "Number of trials necessary to achieve performance stability of selected ground reaction force variables during landing," *Journal of Sports Science & Medicine*, vol. 6, p. 126–134, -3-01 2007.

[24] Scott, "Answer to "at what point do users notice the gradient change in colors?"," -07-11 2017.

[25] J. C. Castro-Alonso and J. Sweller, "The modality effect of cognitive load theory," *Advances in Intelligent Systems and Computing*, vol. 963, pp. 75–84, 06 2019.

[26] U. of Rochester, "Video games lead to faster decisions that are no less accurate," September 13 2010.

[27] U. of Rochester, "Video games lead to faster decisions that are no less accurate," September 13 2010.

[28] "Understanding qq plots | uva library,"

[29] "Normality test: What is normal distribution? methods of assessing normality | editage," -09-27 2022.

[30]

[31] JMP Statistical Discovery, "The T-distribution." https://www.jmp.com/en/statistics-knowledge-portal/inferential-statistics/probability-distributions/t-distribution, 2026. JMP Statistics Knowledge Portal.

[32] R. Bevans, "T-distribution: What it is and how to use it," -08-28 2020.

[33] M. Delacre, C. Leys, Y. L. Mora, and D. Lakens, "Correction: Why Psychologists Should by Default Use Welch's T-test instead of Student's T-test," *International Review of Social Psychology*, vol. 35, no. 1, 2022.

[34] S. Rays, "Introduction to Statistical Testing in R Part 5— Correlation & Regression." https://medium.com/@serurays/introduction-to-statistical-testing-in-r-part-5-correlation-regress Nov 2024. Medium.

[35] Laerd Statistics, "Spearman's Rank-Order Correlation." https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php, 2018. Statistical Guide.

[36] "The decision lab - behavioral science, applied.."

[37] J. Frost, "5 ways to find outliers in your data." https://statisticsbyjim.com/basics/outliers/, Oct 2019. Statistics by Jim.

[38] "Significant digits tutorial | physics."

[39] Kaggle, "Normality Checks: The Shapiro-Wilk Test - a Comprehensive Overview." https://www.kaggle.com/discussions/general/432129, 2024. Kaggle Discussion.

[40] P. Dave, "Managing Sensitivity to Outliers in Regression." https://medium.com/@parth.dave.ca/managing-sensitivity-to-outliers-in-regression-b0c576649ad7, Jun 2024. Medium.

[41] F. Bérard, G. Casiez, and D. Vogel, "On the Limits of the Human Motor Control Precision: The Search for a Device's Human Resolution," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, (New York, NY, USA), pp. 1073–1082, ACM, 2011.

6 Appendices

6.1 Appendix A: Participant Form & Post-Test Survey

A link to the participant form can be found below: https://docs.google.com/forms/d/e/1FAIpQLSczcx viewform?usp=header

A link to the post-test survey questions that were asked to all participants immediately after their completion of the test: https://docs.google.com/document/d/1ufc6iUMie3V6-SwcyP_nKV-Rvkxq9UYIJZ_ CrklxQqg/edit?usp=sharing

6.2 Appendix B: Script

A link to the script that each participant was shown at the beginning of the test: https://docs.google.com/document/d/1gJg4LpQOil-J3uRqdJ1OZoQ3zaqC-cKQMq9CqJJkGKc/edit?usp=sharing Note that in between, starts and italicized text mark times when the test website was used.

6.3 Appendix C: UI Appearance for Modalities

Below are some screenshots of key elements of our website used to conduct all the testing. Note that consistently throughout, the dark purple line represents the line the participant must trace with their cursor, while the dark green line shows the actual line the participant's cursor has made.

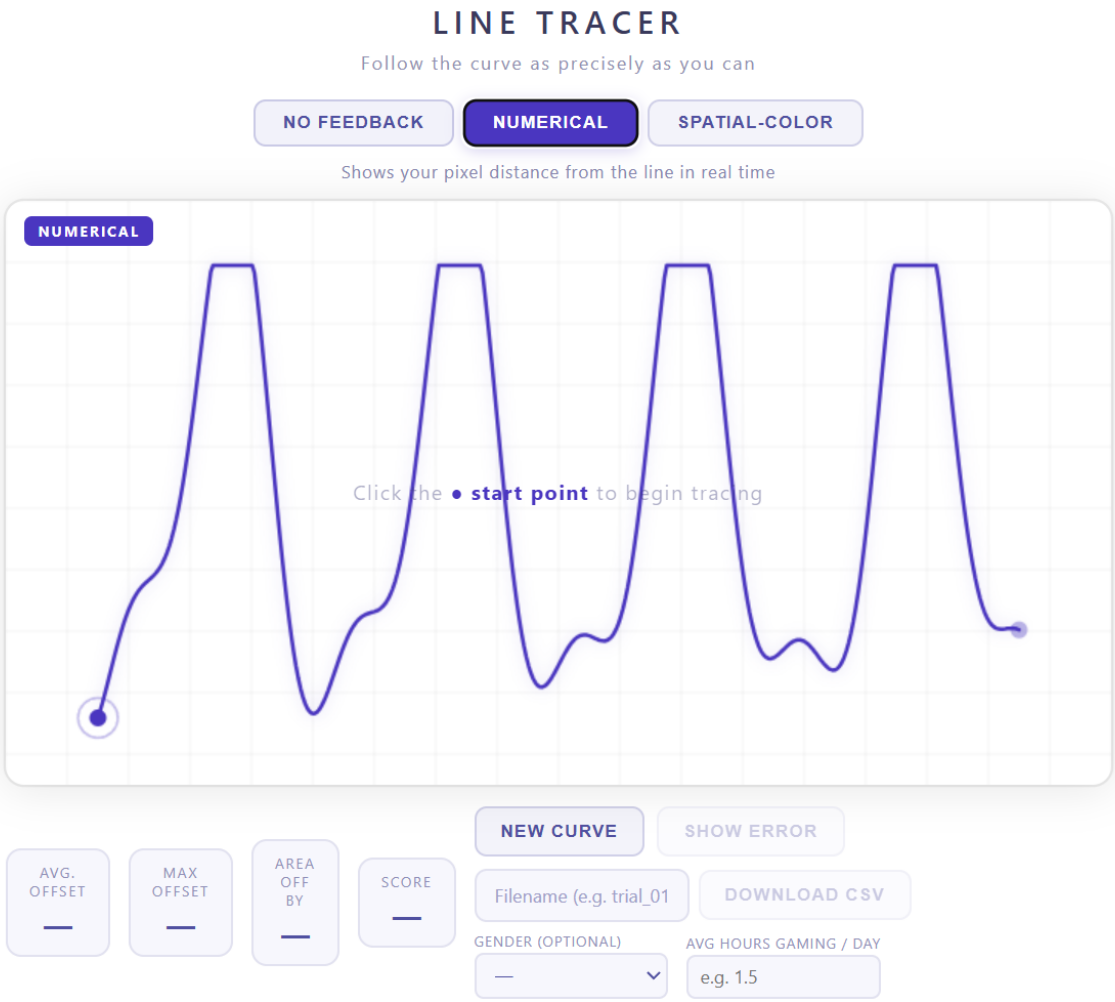

Figure 4: Screenshot of the website when user is initially beginning testing, prior to starting the actual line-tracing test. The curve is shown in dark purple.

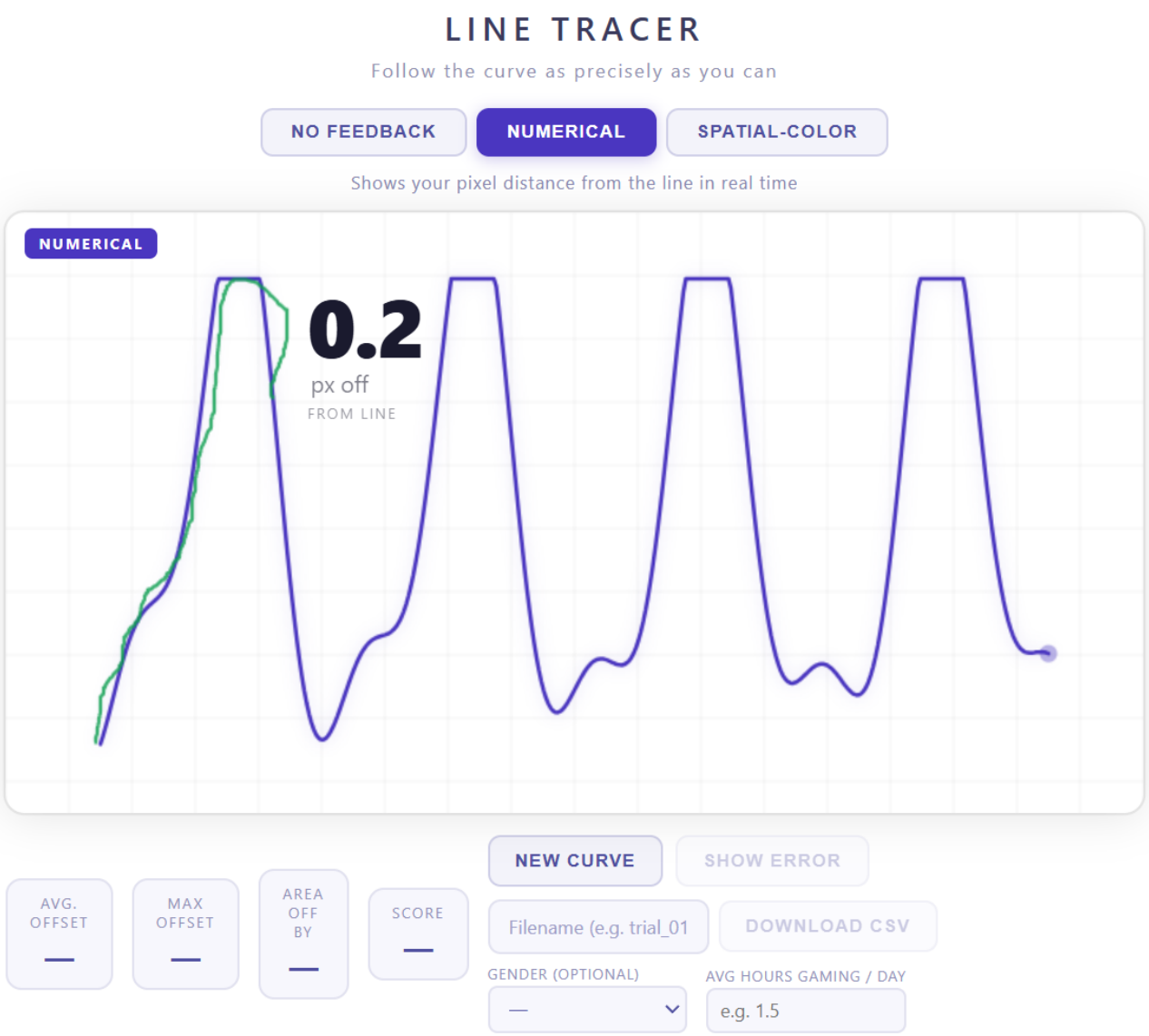

Figure 5: Screenshot of an in-progress numerical test, where the number of pixels the participant is away from the curve is shown in black next to their cursor. The line traced out by the participant can be seen in green.

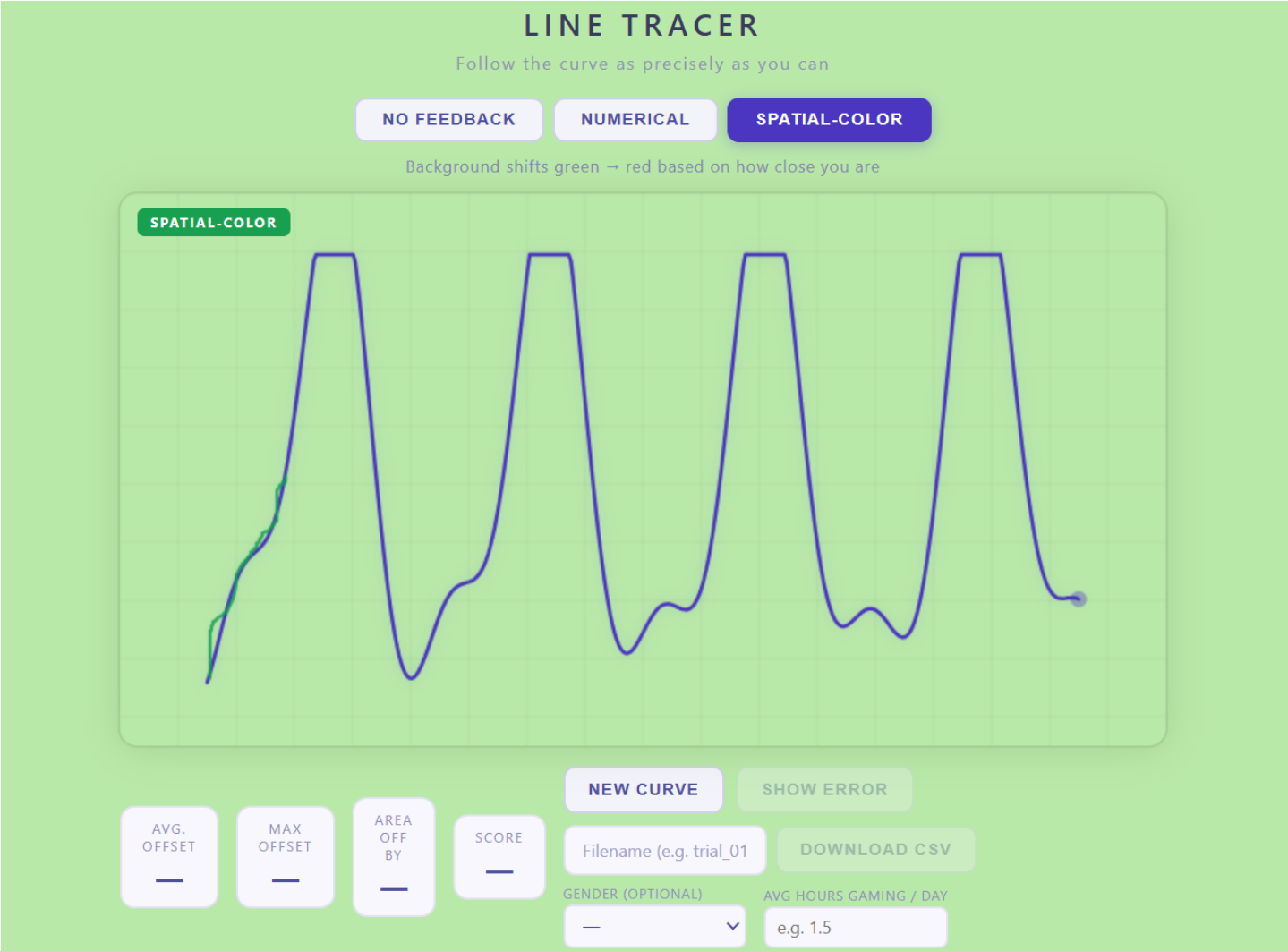

Figure 6: Screenshot of an in-progress spatial test, where the background is currently green as the participant's cursor is extremely close to the line. The line the participant has traced using their cursor can be seen in dark green.

Figure 7: Screenshot of an in-progress spatial test, where the background is currently red as the participant's cursor is far from the line as indicated by the dark green line. Note that the dark green line is the line the participant has traced using their cursor.

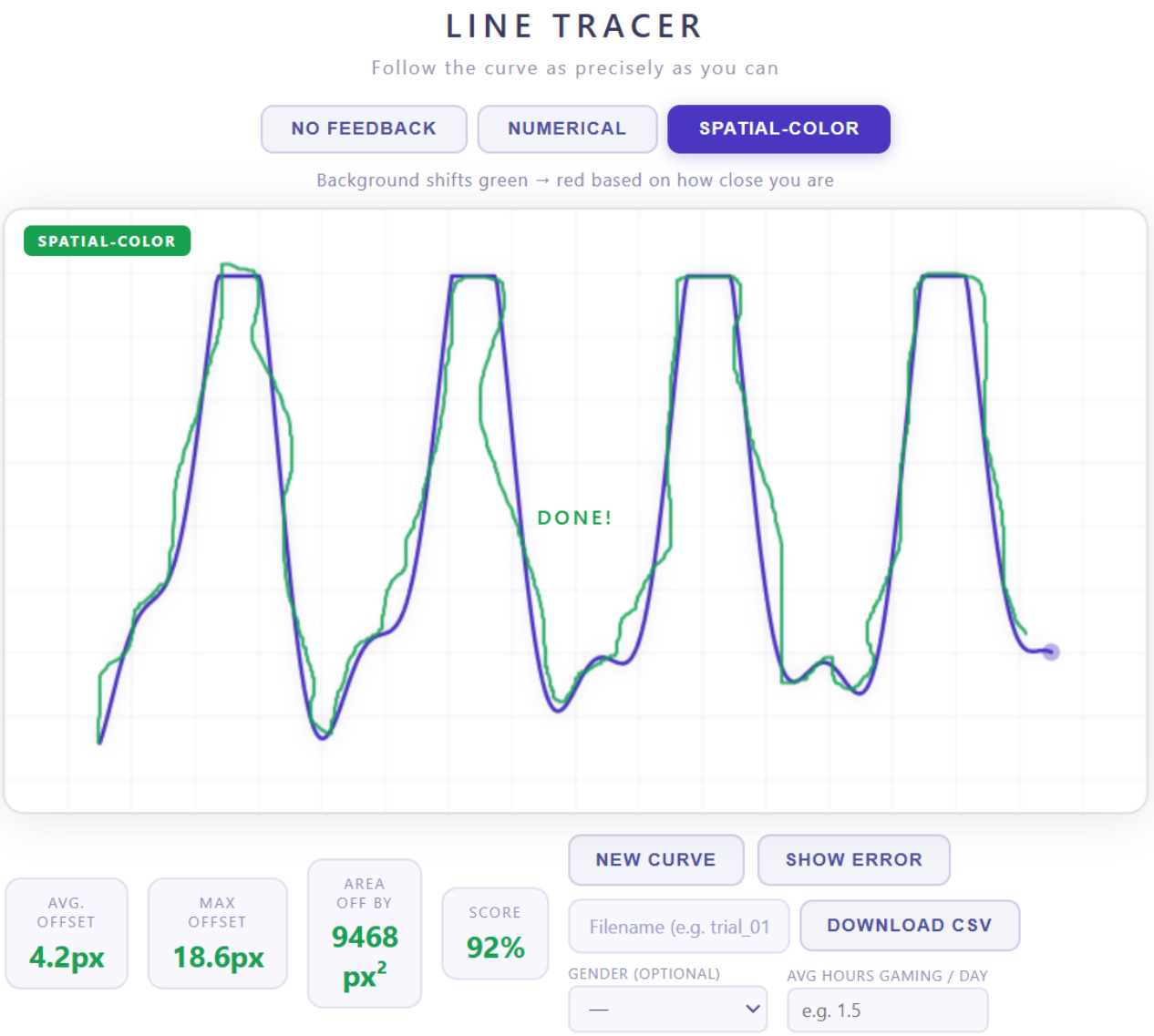

Figure 8: Screenshot of the completed test, with the dark green line indicating the line the participant's cursor has traced out, while the dark purple line is the line the participant had to trace. The download csv button can be seen, where data was downloaded and processed using code.

6.4 Appendix D: Normality Assumption Checks

Figure 9 shows the quantile-quantile plot used to qualitatively assess normality in the sample.

Table 6 shows the output parameters for the Shapiro-Wilk test for the entire $N = 32$ dataset.

Condition	W	p
Numerical Runtime	0.927	0.033
Numerical Accuracy	0.955	0.195
Spatial-colour Runtime	0.938	0.065
Spatial-colour Accuracy	0.894	0.004

Table 6: Shapiro–Wilk test on pooled outcomes ($N = 32$ per cell).

Figure 10 shows the quantile-quantile plot used to qualitatively assess normality in the outlier analysis.

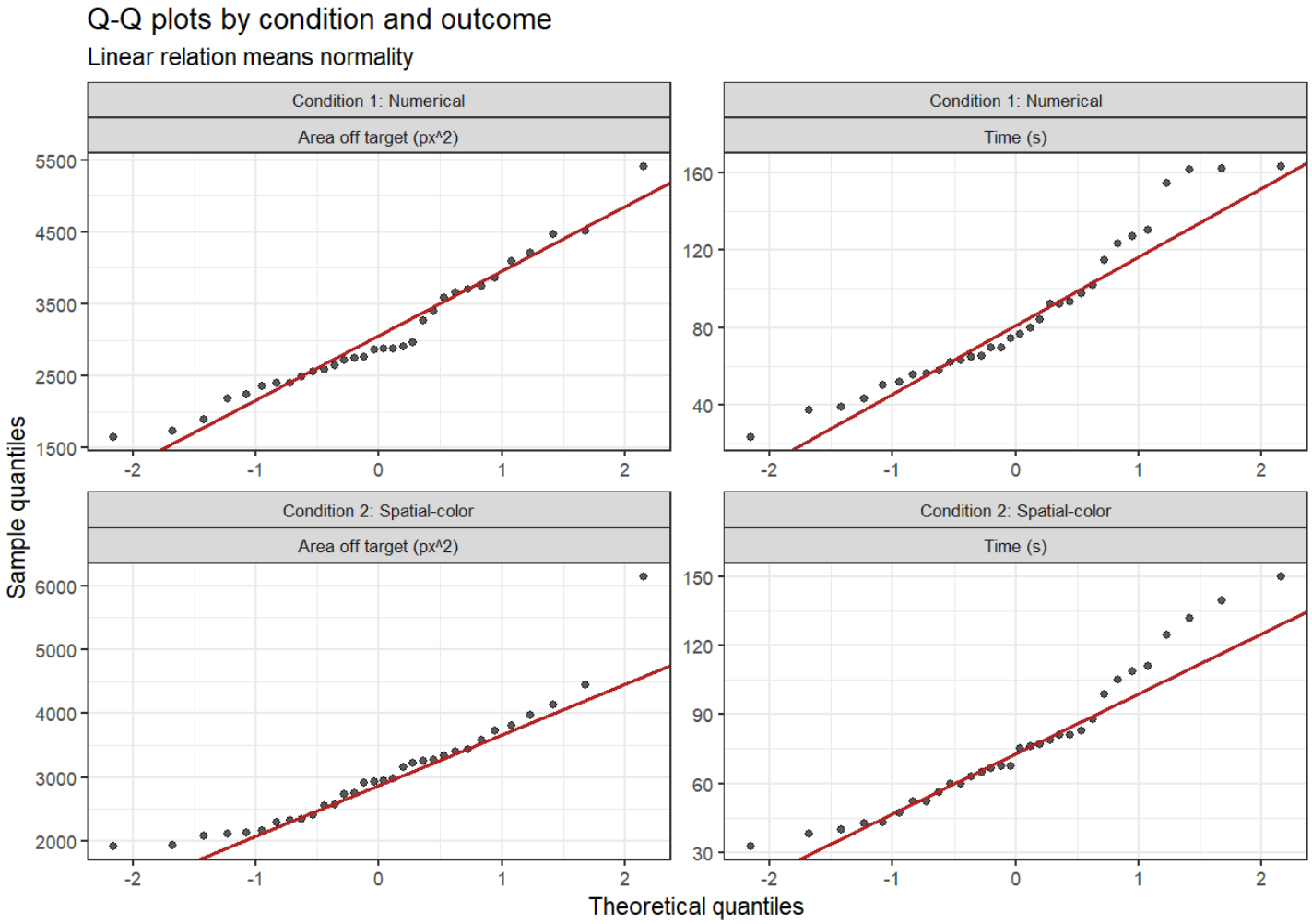

Figure 9: QQ-plots showcasing normality by condition and outcomes, where Conditions 1, 2, 3 and 4 are Numerical Area off target px^2, Numerical Runtime, Spatial-Colour Area off Target and Spatial-Colour Time, respectively.

Table 7 includes the Shapiro-Wilk values for the specificity (outlier analysis).

Condition	W	p
Numerical Runtime	0.956	0.284
Numerical Accuracy	0.956	0.271
Spatial-colour Runtime	0.924	0.043
Spatial-colour Accuracy	0.970	0.580

Table 7: Shapiro-Wilk after excluding outliers (restricted sample, $N = 28$).

6.5 Appendix E: Covariate Figures and Tables

6.5.1 Gender Covariate

The QQ plot for gender is shown in Figure 11. All lines are approximately normal, and by inspection are linear.

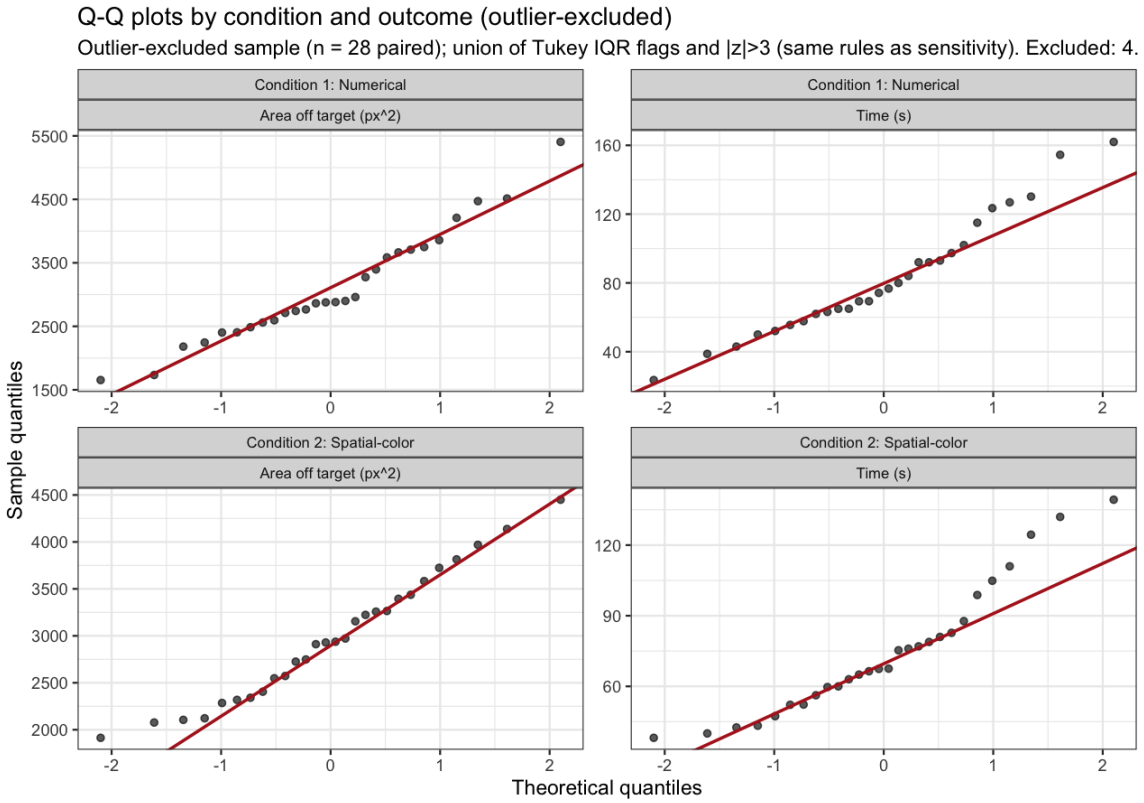

Figure 10: QQ-plots showcasing normality for outliers by condition and outcomes, where Conditions 1, 2, 3 and 4 are Numerical Area off target px^2, Numerical Runtime, Spatial-Colour Area off Target and Spatial-Colour Time, respectively.

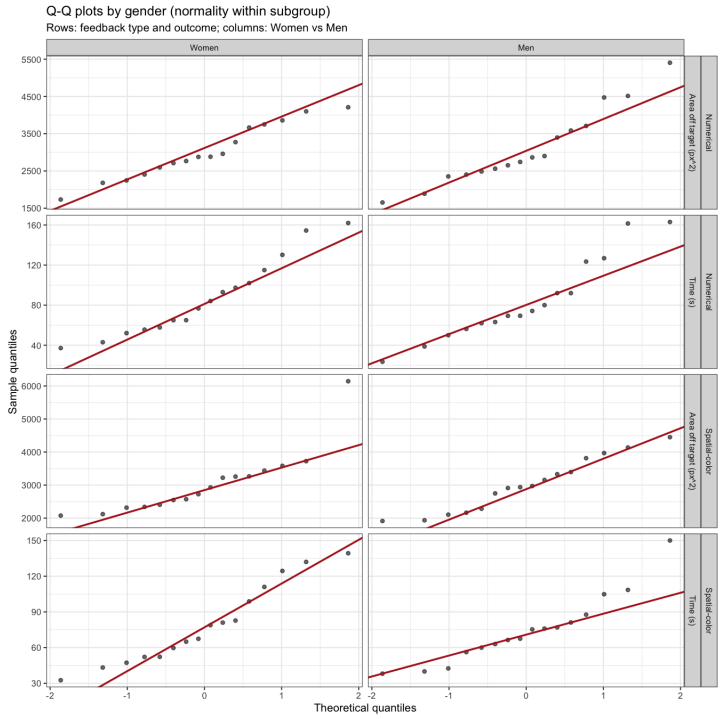

Figure 11: Quantile-quantile plot for performance per gender on speed and accuracy for each modality.

The Shapiro-Wilk results for gender is shown in Figure 8.

Figure 12 shows the box plot for test completion time and accuracy per gender.

Group	Condition	W	p
Females	Numerical Runtime	0.933	0.271
	Numerical Accuracy	0.960	0.655
	Spatial-colour Runtime	0.934	0.285
	Spatial-colour Accuracy	0.779	0.001
Men	Numerical Runtime	0.921	0.173
	Numerical Accuracy	0.930	0.248
	Spatial-colour Runtime	0.913	0.129
	Spatial-colour Accuracy	0.949	0.467

Table 8: Shapiro–Wilk by gender and feedback mode ($N = 16$ per group per stratum).

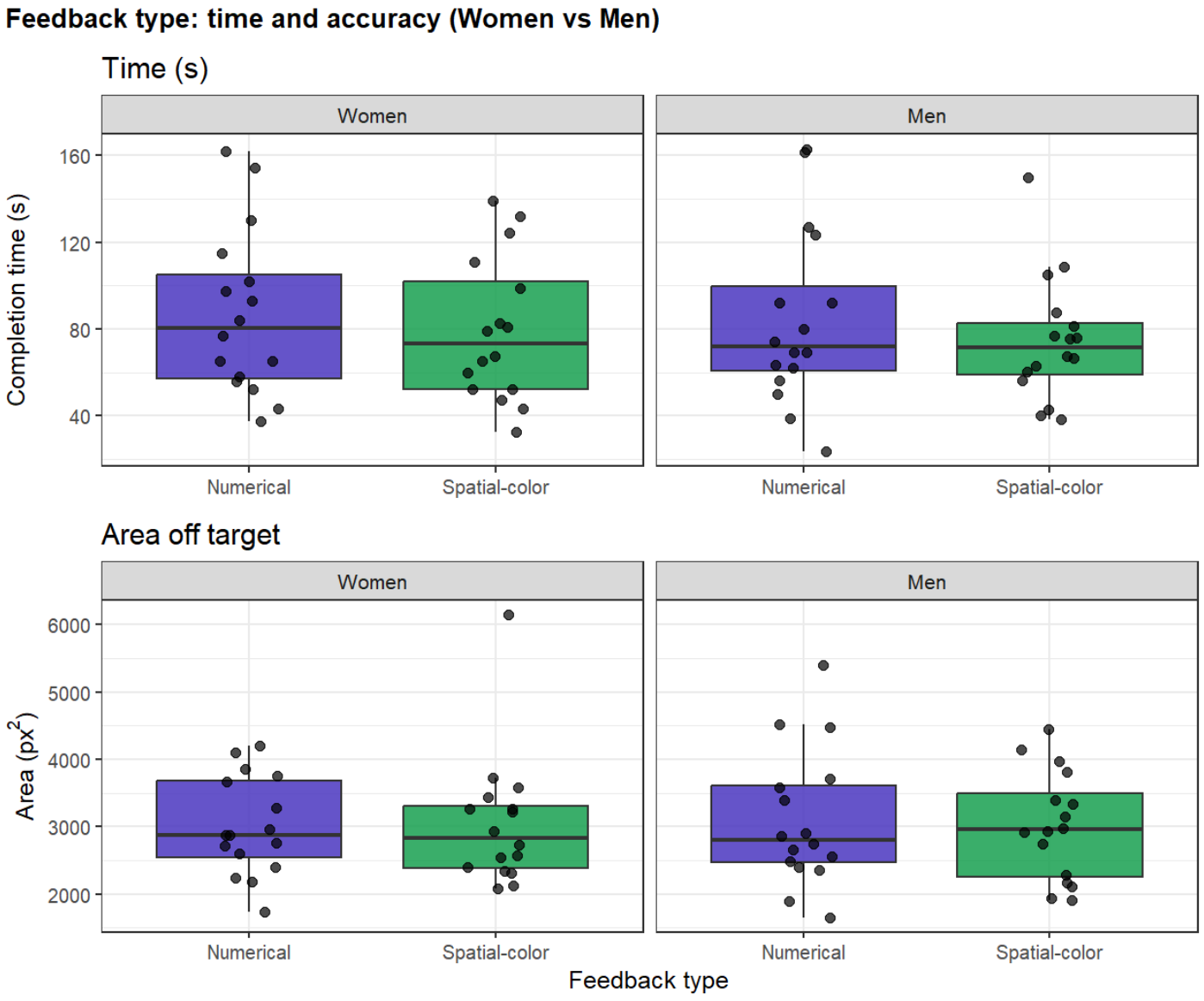

Figure 12: Box Plots for test completion time in seconds (top) and performance accuracy using the area of pixels off target line (bottom), split by gender.

Table 9 shows the results from the t test done per gender, and the outcomes for accuracy and time for both modalities.

Table 9: Welch two-sample t-tests for gender (Females vs. men). H_0: equal means. Two-sided p-values.

Outcome (subset)	t **(df)**	p
Time, numerical (A)	0.20 (29.9)	.842
Time, spatial-colour (B)	0.42 (29.4)	.678
Area, numerical	−0.27 (27.2)	.787
Area, spatial-colour	0.09 (28.9)	.930
Paired diff. time (A. − B)	−0.20 (25.2)	.841
Paired diff. area (A. − B.)	−0.52 (29.2)	.609

6.5.2 Gaming Covariate

The QQ plot for gaming is shown in Figure 13. All lines are approximately normal, and by inspection are linear.

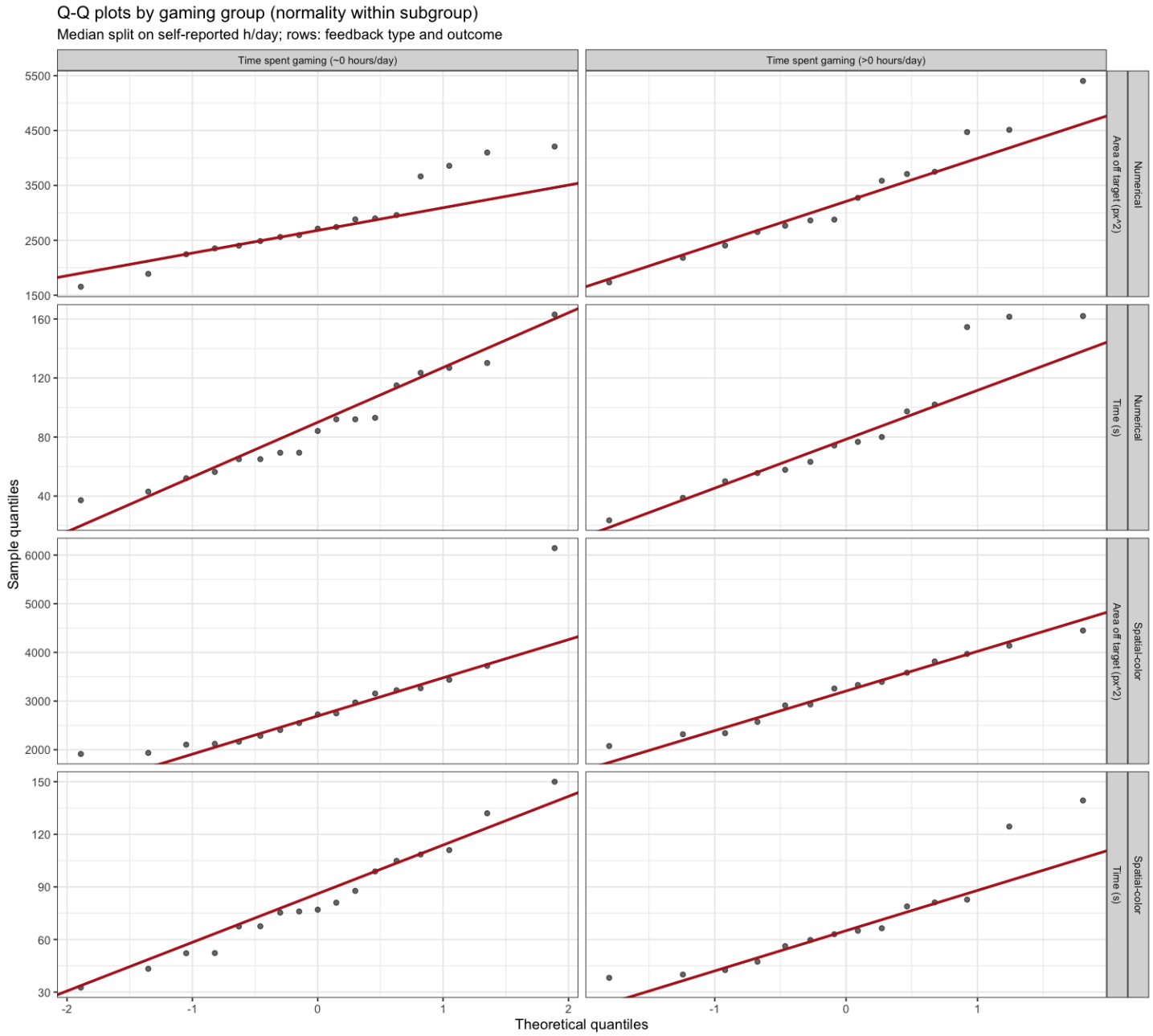

Figure 13: Quantile-quantile plot for performance based on people who game for more than 0 hours and those above 0 hours on speed and accuracy for each modality.

The Shapiro-Wilk results for gender is shown in Figure 10.

Figure 14 shows the box plot for test completion time and accuracy for gaming.

Group	Condition	W	p
~0 h/day	Numerical Runtime	0.951	0.474
	Numerical Accuracy	0.929	0.206
	Spatial-colour Runtime	0.974	0.889
	Spatial-colour Accuracy	0.778	0.001
>0 h/day	Numerical Runtime	0.892	0.086
	Numerical Accuracy	0.967	0.827
	Spatial-colour Runtime	0.863	0.034
	Spatial-colour Accuracy	0.969	0.864

Table 10: Shapiro–Wilk by gaming-time split and feedback mode.

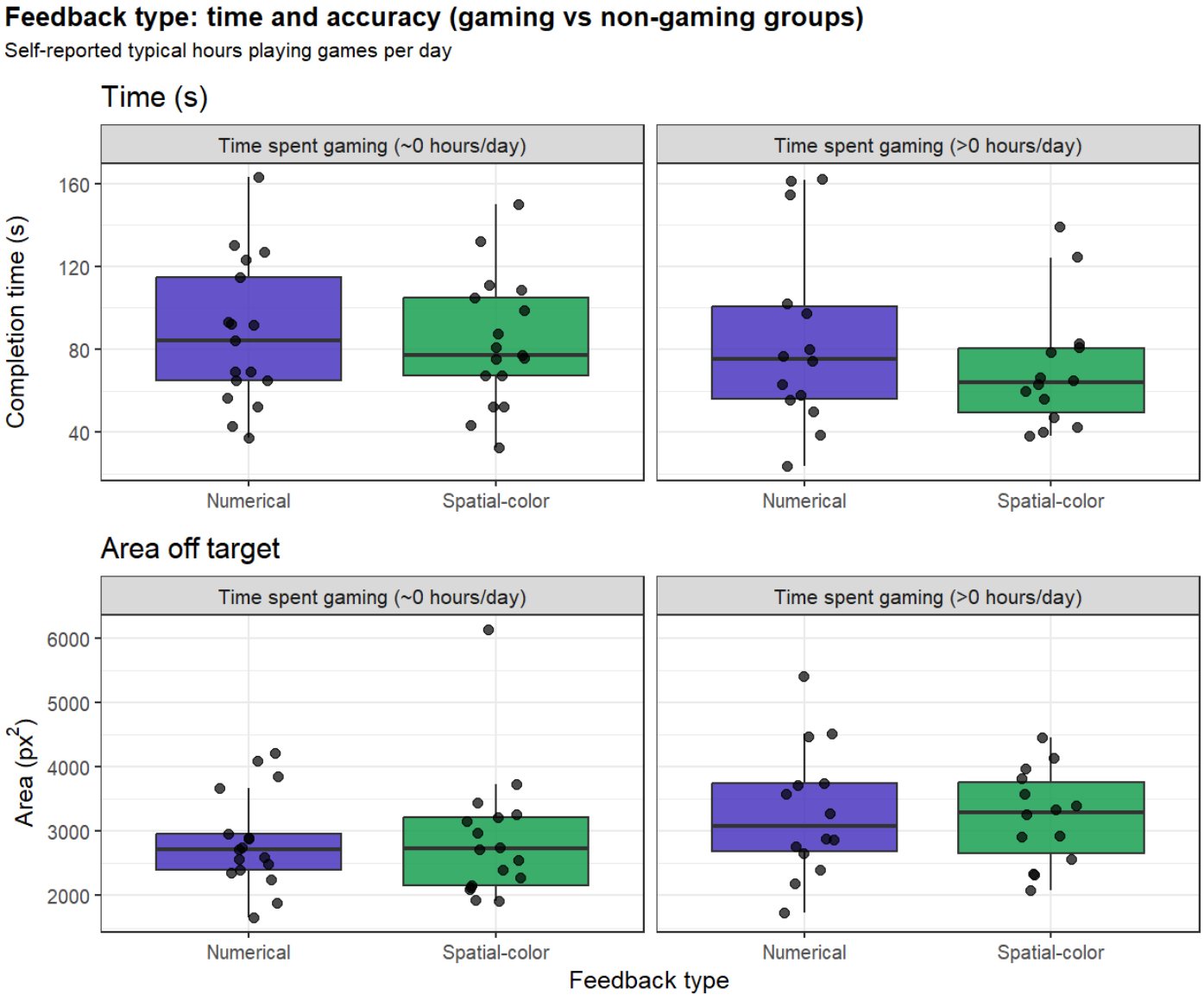

Figure 14: Box Plots for time taken to complete the test in seconds (top) and performance accuracy as measured by the area of pixels off target line (bottom), split by the amount of time participants spent gaming per day

Table 11 shows the results from the t test done for gaming, and the outcomes for accuracy and time for both modalities.

Table 11: Welch two-sample t-tests for gaming groups. H_0: equal means. Two-sided p-values.

Outcome (subset)	t (df)	p
Time, numerical	0.09 (24.2)	.927
Time, spatial-colour	1.18 (28.3)	.248
Area, numerical	−1.43 (23.1)	.166
Area, spatial-colour	−1.10 (28.6)	.279
Paired diff. time (num. − spa.)	−1.30 (27.1)	.205
Paired diff. area (num. − spa.)	−0.53 (29.0)	.603

6.6 Appendix F: Key Code Snippets

Below are some of the code we used to analyze our data in R.

6.6.1 *Appendix F.2: Shapiro-Wilk (W) Code*

```r
shapiro_by_group <- assump %>%
  group_by(condition, outcome) %>%
  group_modify(~ {
    v <- .x$value
    n <- length(v)
    statistic_W <- NA_real_
    p_value <- NA_real_
    statistic_D <- NA_real_
    p_value_ks <- NA_real_
    if (n >= 3L && n <= 5000L) {
      sw <- stats::shapiro.test(v)
      statistic_W <- unname(sw$statistic)
      p_value <- sw$p.value
    }
    if (n >= 4L && n <= 5000L) {
      lil <- nortest::lillie.test(v)
      statistic_D <- unname(lil$statistic)
      p_value_ks <- lil$p.value
    }
    tibble::tibble(
      n = n,
      statistic_W = statistic_W,
      p_value = p_value,
      statistic_D = statistic_D,
      p_value_ks = p_value_ks
    )
  }) %>%
  ungroup()
print(shapiro_by_group, n = Inf)
cat("\n")
```

Figure 15: Example implementation of Shapiro-Wilk (W) calculations. This code can be found in mie286_analysis_pipeline.R

The object `shapiro_by_group` stores normality test results for each condition-outcome group. Within `group_modify()`, the Shapiro-Wilk test (`stats::shapiro.test()`) was applied for groups where `n` fell between 3 and 5000. The test returned a W statistic and p-value, with `NA` assigned where sample size requirements were not met.

6.6.2 Appendix F.2: Quantile-Quantile (QQ) Plots Code

```r
qq_pts <- assump %>%
  group_by(condition, outcome) %>%
  group_modify(~ {
    v <- sort(as.numeric(stats::na.omit(.x$value)))
    n <- length(v)
    if (n < 2L) return(tibble::tibble(theoretical = NA_real_, sample = NA_real_))
    tibble::tibble(theoretical = stats::qnorm(stats::ppoints(n)), sample = v)
  }) %>%
  ungroup() %>%
  filter(!is.na(theoretical))
```

Figure 16: Example implementation of QQ plot. This code can be found in mie286_analysis_pipeline.R

The object `qq\_pts` stores the data used to construct the QQ plot by pairing sample values with their corresponding theoretical normal quantiles. For each group, the vector v represents the sorted sample data after removing missing values while the variable n denotes the number of observations in this vector. Theoretical quantiles were computed using `stats::ppoints()` and then converted to standard normal quantiles using `stats::qnorm()`.

6.6.3 *Appendix F.3: One-Tailed Test Code*

```r
cat("Paired t-tests (two-tailed)\n")
tt_time <- t.test(
  paired_complete$`duration_sec___numerical`,
  paired_complete$`duration_sec___spatial-color`,
  paired = TRUE
)
print(tt_time)
tt_area <- t.test(
  paired_complete$`area_off_px2___numerical`,
  paired_complete$`area_off_px2___spatial-color`,
  paired = TRUE
)
print(tt_area)
cat("\n")
```

Figure 17: Example implementation of one-tailed directional paired *t*-tests. This code can be found in mie286_analysis_pipeline.R

Paired *t*-tests were conducted using `t.test()` with `paired = TRUE` and explicit directional alternatives. For runtime (**Hypothesis 2.1**), we used `alternative = "greater"` on `numerical - spatial-color` to test whether spatial-colour feedback is faster. For accuracy (**Hypothesis 2.2**), we used `alternative = "less"` on `numerical - spatial-color` to test whether numerical feedback is more accurate. Results were printed directly for each outcome.

6.6.4 *Appendix F.4: Welch's t-test Code*

```r
if (length(tab_g) >= 2L && min(tab_g) >= 2L) {
  cat("Mode: ", md, " — Shapiro–Wilk normality within gender (duration, s)\n", sep = "")
  shapiro_or_skip_subgroup(dg$duration_sec[dg$gender_f == "Women"], "  Women")
  shapiro_or_skip_subgroup(dg$duration_sec[dg$gender_f == "Men"], "  Men")
  cat("Mode: ", md, " — Welch t duration (s)\n", sep = "")
  mie286_print_htest_no_ci(stats::t.test(duration_sec ~ gender_f, data = dg, var.equal = FALSE))
  cat("Mode: ", md, " — Shapiro–Wilk normality within gender (area off target)\n", sep = "")
  shapiro_or_skip_subgroup(dg$area_off_px2[dg$gender_f == "Women"], "  Women")
  shapiro_or_skip_subgroup(dg$area_off_px2[dg$gender_f == "Men"], "  Men")
  cat("Mode: ", md, " — Welch t area (px^2)\n", sep = "")
  mie286_print_htest_no_ci(stats::t.test(area_off_px2 ~ gender_f, data = dg, var.equal = FALSE))
  cat("\n")
```

Figure 18: Example implementation of Welch's t-test test calculation. This code can be found in mie286_analysis_pipeline.R

A Welch two-sample *t*-test was conducted using `stats::t.test()` with `var.equal = FALSE`, comparing group means using formula syntax (`outcome ~ group`). Setting `var.equal = FALSE` instructs R to use separate sample variances for each group and apply the Welch–Satterthwaite approximation to adjust the degrees of freedom, accounting for heteroscedasticity between groups. The specific outcome and grouping variables may differ across applications of this test.

6.6.5 *Appendix F.5: Pearson's Coefficient Code*

```r
cat("Pearson: duration vs area (Numerical)\n")
cor_num <- cor.test(act_num$duration_sec, act_num$area_off_px2, method = "pearson")
print(cor_num)
```

Figure 19: Example implementation of the Pearson's coefficient calculation. This code can be found in mie286_analysis_pipeline.R

A Pearson correlation coefficient was computed using `cor.test()` with method = "pearson" to assess the linear relationship between two continuous variables, here `duration_sec` and `area_off_px2`. The specific variables and conditions may differ across applications of this test.

6.6.6 *Appendix F.6: Z-Score and IQR Code*

```r
col_z_extreme <- function(v, zmax = 3) {
  m <- mean(v, na.rm = TRUE)
  s <- stats::sd(v, na.rm = TRUE)
  if (!is.finite(s) || s < 1e-12) {
    return(rep(FALSE, length(v)))
  }
  abs((v - m) / s) > zmax
}
```

Figure 20: Implementation of Z-score outlier detection. This code can be found in mie286_outlier_rules.R

Outliers were flagged using a z-score function `col_z_extreme()`, which computes the mean and standard deviation of a vector and flags any observation where the absolute z-score exceeds a threshold, defaulting to `zmax = 3`. If the standard deviation is non-finite or near-zero, no observations are flagged to avoid division errors.

```r
col_iqr_extreme <- function(v) {
  q <- stats::quantile(v, c(0.25, 0.75), na.rm = TRUE, names = FALSE)
  iqr <- q[2] - q[1]
  if (!is.finite(iqr) || iqr <= 0) {
    return(rep(FALSE, length(v)))
  }
  lo <- q[1] - 1.5 * iqr
  hi <- q[2] + 1.5 * iqr
  v < lo | v > hi
}
```

Figure 21: Implementation of IQR outlier detection. This code can be found in mie286_outlier_rules.R

Outliers were also flagged using an IQR-based function `col\_iqr\_extreme()`, which computes the 25th and 75th percentiles and flags any observation falling below Q1 - 1.5 × IQR or above Q3 + 1.5 × IQR,

following the standard Tukey method. If the IQR is non-finite or zero, no observations are flagged.

6.6.7 Appendix F.7: Mean and Standard Deviation Code

```r
curve_df <- assump %>%
  group_by(condition, outcome) %>%
  group_modify(~ {
    v <- .x$value
    m <- mean(v, na.rm = TRUE)
    s <- stats::sd(v, na.rm = TRUE)
    if (!is.finite(s) || s < 1e-9) s <- 1e-6
    rg <- range(v, na.rm = TRUE)
    pad <- max(diff(rg) * 0.25, 1, na.rm = TRUE)
    xs <- seq(rg[1] - pad, rg[2] + pad, length.out = 200)
    tibble::tibble(x = xs, y = stats::dnorm(xs, m, s))
  }) %>%
  ungroup()
```

Figure 22: Example implementation of mean and standard deviation calculation by group. This code can be found in mie286_analysis_pipeline.R

For each condition-outcome group, observed values were stored as a vector v. The mean and standard deviation were computed directly on v using `mean()` and `stats::sd()` respectively. A near-zero or non-finite standard deviation is replaced with a small fallback value to avoid errors in downstream calculations.

6.7 Appendix G: Participant Post-Test Results to Questions

The results here refer to the questions asked in Appendix B after the test.

Figure 23 uses the participant feedback given after the test and quantifies them into distinct categories, being their modality preference, the common theme we saw in their numerical feedback comments, and if they did or did not learn the curve.

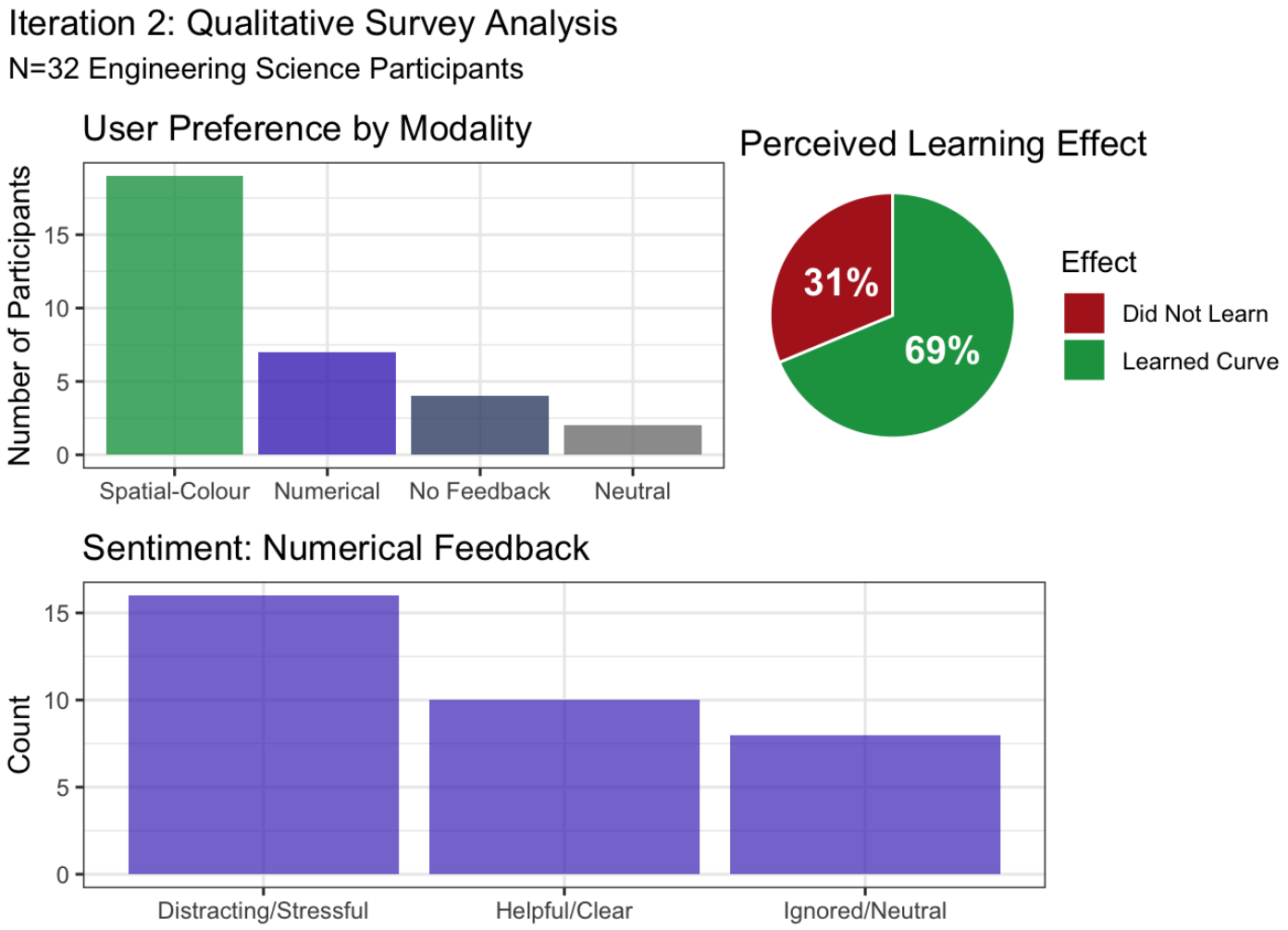

Figure 23: Three different plots are shown: the top right is a bar chart of the user preference per modality (spatial-color, numerical, no feedback, or no opinion), the top right is a pie chart comparing how many participants learned and did not learn the curve, and the bottom bar chart reflects the user's common sentiments on numerical feedback.