Optimizing User Engagement:
Experimental Design to Minimize Decision Paralysis on Webflix

Amadeo Cabanela, John Green, Ricky Miura

MSDS 629
Experiments in Data Science, Intersession 2025

Professor Nathaniel Stevens

EXECUTIVE SUMMARY

In today's world of endless streaming options, Webflix users often face decision paralysis, leading to longer browsing times and increased frustration. This project aims to optimize the "Today's Top Picks for You" row on the Webflix homepage to alleviate this issue by fine-tuning four key design factors: Match Score, Preview Length, Preview Type, and Tile Size. Our series of experiments, which included Factorial Screening, Response Surface Methodology, and Grid Search, revealed that Match Score and Preview Length were significant factors, motivating us to prioritize them for further exploration. As a result, we determined the optimum location to be at a Match Score of 70 percent, Preview Length of 70 seconds, Preview Type as teaser/trailer (TT), and Tile Size of 0.2 (default value), resulting in a minimized average browsing time of 10.497 minutes with a 95% confidence interval of (10.46, 10.54) minutes.

INTRODUCTION

When selecting content on the Webflix streaming service, users often face an overwhelming number of choices, leading to a phenomenon known as decision paralysis. For example, a user might sit down for dinner intending to relax with a movie or show but instead spend the time endlessly scrolling. By the time their meal is cold, no decision has been made, resulting in frustration and disengagement. Webflix tries to help users decide on something suitable to watch by providing personalized recommendations with the “Today’s Top Picks for You” row on the homepage.

Our goal is to optimize this row’s design to address the problem of decision paralysis and minimize average browsing time (in minutes), which is our metric of interest. We focus on examining the impact of four key design factors: Tile Size (tile height-to-screen ratio, default: 0.2), Match Score (enjoyment prediction in percent, default: 95), Preview Length (preview duration in seconds, default: 75), and Preview Type (preview autoplay type, default: TT for teaser/trailer). Our experiments began with Factor Screening to identify significant factors. After determining that Tile Size was not significant and the optimal Preview Type was TT, we focused on optimizing Match Score and Preview Length using Response Surface Methodology and Grid Search. This report outlines our experiments, the reasoning behind our design choices, results, and statistical analysis, offering practical recommendations for enhancing the Webflix browsing experience.

EXPERIMENTS

Experiment 1: *Factor Screening Using a 2^4 Factorial Experiment*

We began with a 2^4 full factorial experiment for factor screening to answer the question: "Which among the four design factors significantly influences average browsing time?" The objective was to efficiently narrow down the search space, saving time and resources, so we could focus on studying the most impactful factors in subsequent experiments.

Each factor was tested at two extreme levels of their ranges to maximize the detection of significant effects (Table 1). Levels were encoded as −1 (low) and +1 (high). We did a power analysis to determine the appropriate sample size. We chose a moderate effect size of $d = 0.4$ (consistent with Cohen's Guidelines for behavioral studies), the standard significance level $\alpha = 0.05$, and power $1-\beta = 0.8$. Although the power calculation yielded 100 units, we assigned 250 experimental units to each condition to enhance the robustness of our findings for 4,000 total observations across all $2^4 = 16$ conditions.

We collected the data, then fitted a first-order linear regression model with interaction terms. The −1/+1 encoding creates an orthogonal design, allowing us to interpret the t-tests from the model summary and draw conclusions to identify significant factors ($p \leq 0.05$). Tile Size showed no significant main effects or interactions. Both Match Score and Preview Length had significant main effects and a significant interaction as reflected in the interaction plot (Figure 1) which shows a clear departure from parallel lines. When Match Score was 0, Preview Length had little impact on average browsing time. However, when Match Score was 100, increasing Preview Length from 30 to 120 notably increased average browsing time. We therefore prioritized higher match scores in future experiments. Preview Type had a significant main effect but not interactions; the high level (AC) had a positive coefficient, indicating increased browsing time (which is not optimal) compared to the reference (TT). Based on these findings, we fixed Tile Size at its default 0.2 and Preview Type at TT for all succeeding experiments to focus future exploration on Match Score and Preview Length.

Factor	Low (–)	High (+)
Tile Size	0.1	0.5
Match Score	0	100
Preview Length	30	120
Preview Type	TT	AC

Table 1

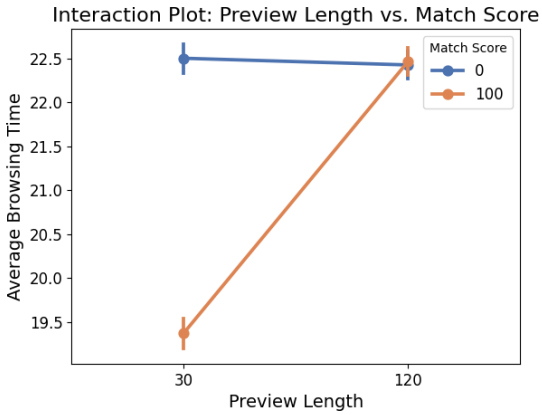

Figure 1

Experiment 2: *Narrowing Optimal Ranges for Preview Length and Match Score*

Next, we further explored our two main factors. The question we aimed to answer was, "What range of values in Preview Length and Match Score corresponds to lower average browsing times?" Our objective was to narrow the optimal range for these two factors to guide response optimization in later experiments. We explored three levels, dispersed evenly throughout the parameter space, combined with the two extremes tested previously. This resulted in five levels for Preview Length ({30, 55, 75, 95, 120}) and Match Score ({0, 25, 50, 75, 100}), yielding $5^2 = 25$ experimental conditions. We fixed Tile Size at 0.2 and Preview Type at TT as established in the previous experiment. Repeating the power analysis yielded similar results as before, so we again assigned 250 units per condition. Since we already had data for the extreme levels, we collected an additional 2,250 observations for the three middle levels.

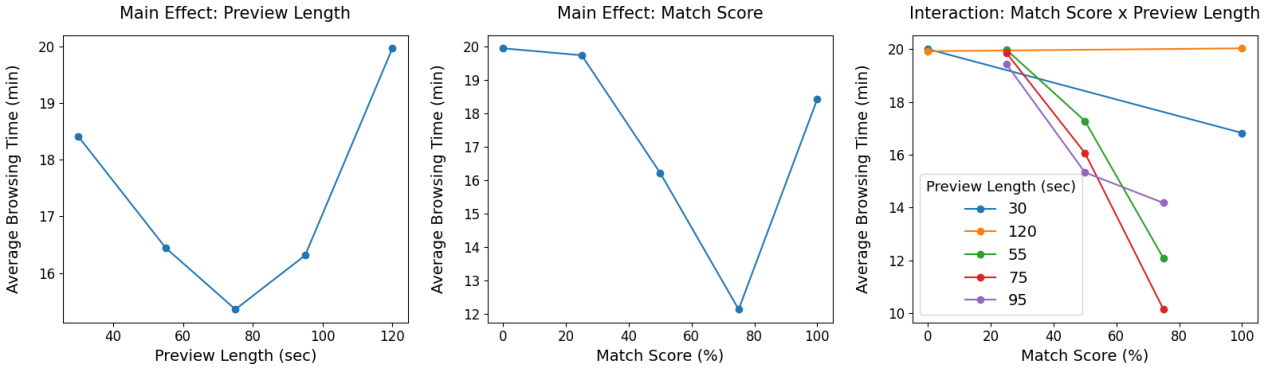

Figure 2

Based on the main effect plots in Figure 2, we hypothesized an optimal range for these two factors, indicated by the valleys in the main effect plots. We conducted two F-tests to evaluate equality among levels with the following hypotheses:

$$(1)\ H_0\colon \mu_{55} = \mu_{75} = \mu_{95}\ \ vs\ \ H_A\colon \mu_j \neq \mu_k \quad \text{for some preview length } j \neq k$$

$$(2)\ H_0\colon \mu_{50} = \mu_{75} = \mu_{100}\ vs\ \ H_A\colon \mu_j \neq \mu_k \quad \text{for some match score } j \neq k$$

The p-values for each experiment were 1.37e-10 and 0.00, respectively. At a 5% significance level, we rejected the null hypotheses in both (1) and (2), concluding that the expected browsing time is not the same across levels. This suggests that both factors significantly influence average browsing time.

We conducted another F-test to further explore the previously identified interaction between Match Score and Preview Length and confirm its statistical significance. Looking at the interaction effect plot in Figure 2, a Match Score of 75 across Preview Lengths {55, 75, 95} had the smallest average browsing times. Thus, we tested the following hypothesis:

$$(3)\ H_0\colon \mu_{55,75} = \mu_{75,75} = \mu_{95,75}\ \ vs\ \ H_A\colon \mu_{j,75} \neq \mu_{k,75} \text{ for some } j \neq k$$

where the subscripts represent the level of Preview Length and Match Score assigned to each unit

The p-value for this experiment was 3.77e-221. At a 5% significance level, we rejected the null hypothesis in (3) and concluded that the expected browsing time is not the same across conditions which suggests that the interaction between Match Score and Preview Length significantly influences average browsing time. Based on the plots and the conclusions of our hypothesis tests, we get a decent idea of the approximate range of where the optimum for each factor is. For Preview Length, it seems to be between 50 and 90. For Match Score, it also looks to be between 50 and 90. Now that we have found these bounds, we can use them as our −1/+1 levels for a Central Composite Design (CCD) experiment.

Experiment 3: *Response Surface Optimization of Browsing Time with Preview Length and Match Score*

After identifying the approximate boundaries for minimizing browsing time – between [50, 90] for Match Score and [50, 90] for Preview Length – our next experiment aimed to answer the question, "What optimal values of Match Score and Preview Length minimize average browsing time?" The objective was to fine-tune these factors to determine their optimal combination. To achieve this, we employed Response Surface Methodology (RSM) with a Central Composite Design (CCD).

A two-factor CCD was selected, using a standard alpha value of $\sqrt{2}$ and 4 × 4 center points to enhance model accuracy and robustness. The encoded factor levels were set at {–1.41, –1, 0, 1, 1.41}. In natural units, the factorial conditions for both Preview Length and Match Score were set to 50 and 90, leading to experimental levels of {40, 50, 70, 90, 100} for both factors. The per-group sample size was determined to be 250 units, based on statistical power calculations discussed previously.

We collected data on 4,000 observations, then we fitted a second-order regression model with first-order interaction terms to characterize the response surface. The model equation is given by:

$$(4)\ \hat{\eta} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \widehat{\beta_{12}} x_1 x_2 + \widehat{\beta_{11}} x_1^2 + \widehat{\beta_{22}} x_2^2$$

where $\hat{\eta}$ is the estimated response surface for browsing time, $\widehat{\beta_0}$ is the intercept, $\widehat{\beta_1}$ and $\widehat{\beta_2}$ are first-order coefficients for Match Score (MS, x_1) and Preview Length (PL, x_2), $\widehat{\beta_{12}}$ is the interaction effect between MS and PL, $\widehat{\beta_{11}}$ and $\widehat{\beta_{22}}$ are second-order coefficients for MS and PL.

We generated the response surface for expected browsing time using this model. Visualizing it as a contour plot (Figure 3) revealed significant curvature, indicating the presence of a non-linear relationship between the factors and expected browsing time, and suggesting the presence of a local minimum. The minimum region was estimated to fall within the range of [60, 85] for Match Score and [50, 90] for Preview Length.

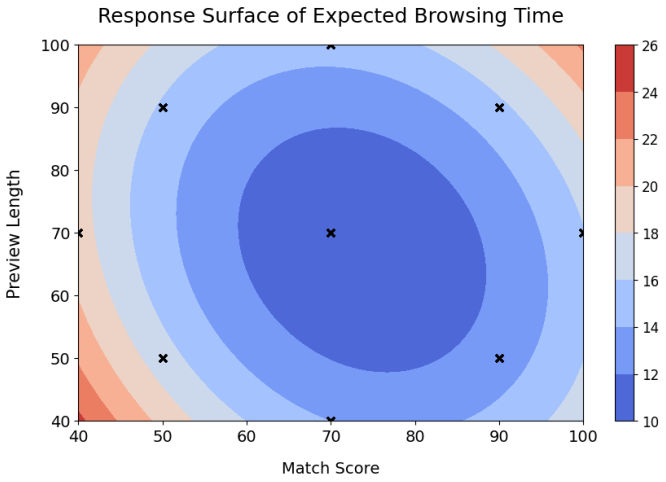

Figure 3

To precisely determine the minimum, we applied the optimization algorithm Broyden-Fletcher-Goldfarb-Shanno (BFGS). The identified optimal conditions were Match Score = 73.71 percent and Preview Length = 67.29 seconds. These conditions led to an estimated minimized average browsing time of 10.78 minutes.

Experiment 4: *Finding the Precise Optimum within Valid Range*

We have identified the optimal combination of factors to minimize average browsing time, but these values were not valid within their defined operational ranges. For example, the optimal Match Score was 73.71, but its valid values are integers between [0, 100]. Also, the optimal Preview Length was 67.29, but valid values fall between [30, 120] in 5-second increments. In this last experiment, we aimed to answer the question, "What integer Match Score and Preview Length divisible by 5 minimizes average browsing time?" The objective was to adjust the optimal Match Score and Preview Length to their nearest valid values so we can confirm their statistical significance and finalize the optimal configuration.

We explored additional points within the contour and found that rounding Preview Length to 65 and Match Score to 75 resulted in a lower predicted average browsing time compared to the previously calculated minimum. This was likely due to some curvature in the area not accounted for by the fitted second-order model. To validate this model, we performed a small grid search in the area immediately surrounding the calculated minimum, testing Preview Lengths {65, 70} and Match Scores {70, 71, 72, 73, 74, 75} on 3,000 data collected observations. Using ANOVA revealed a significant difference in browsing time, with Match Score = 73 and Preview Length = 70 being the best combination in our data. We then tested the significance of these findings through a series of pairwise t-tests using the Bonferroni correction to account for the multiple comparison problem. With Preview Length of 70, the difference in average browsing times across Match Scores was not statistically significant as long as Match Score was between 70 and 75. To conclude, we chose a Preview Length (PL) of 70 and rounded Match Score (MS) to 70, as it represents a more practical choice for future experiment tracking, yielding an average browsing time of 10.59 minutes. Using PL=70 and MS=70, our final quadratic model from Experiment 3 yielded a similar prediction of 10.497 minutes for average browsing time with a 95% confidence interval of (10.46, 10.54).

CONCLUSION

Through a series of carefully designed experiments, we identified the optimal conditions that minimize browsing time on the Webflix platform. Using Response Surface Methodology and subsequent Grid Search, we conclude that a Match Score of 70 percent, Preview Length of 70 seconds, and Preview Type of teaser/trailer (TT) provide the best combination for minimizing browsing time, with Tile Size not relevant (use default 0.2). In our final model, these conditions yield the estimated optimum average browsing time of 10.497 minutes with a 95% confidence interval of (10.46, 10.54) minutes, a notable improvement from the initial baseline. While one might expect a perfect match score to be optimal, our findings suggest that balancing moderate Preview Length with a well-calibrated Match Score effectively addresses decision paralysis, creating a more efficient and engaging selection process for users.

Although our experiments demonstrate statistical significance, several limitations should be considered when interpreting these findings. First, the fitted second-order regression model assumes a smooth, quadratic response surface. While this approach is effective for guiding optimization, it may not fully capture localized irregularities in browsing time behavior. Second, the results are based on specific experimental conditions and may not generalize across all user demographics or content types. Additionally, there may have been nuisance factors that were not accounted for, which could influence the outcomes of the experiments. As such, we must interpret these findings with caution. Despite these limitations, the insights from this study provide a robust foundation for further refinement and practical implementation to improve the Webflix user experience.