The Impact of Interface Design on User Engagement in a Web-Based Personality Quiz

Jiaqi Guo, Yixuan Chen, Jingxi Li, Megha Polavarapu, Safia

4/24/2025

GitHub: https://github.com/meghapola/ADS-Project-3

Introduction & Research Question

In an increasingly digital world—and especially one in which attention spans are decreasing at unprecedented rates—the design and presentation of surveys, quizzes, and other digital data-collection platforms play a crucial role in shaping user experience, engagement, and data quality gathered. With this in mind, our team chose to design a personality quiz that we delivered through a web-based Shiny application. We were curious about how changes in layout, wording, and visual appeal could affect how participants interpret questions, respond to them, and engage with an online quiz at-large.

We designed two versions of the same 20-question personality quiz, which we refer to as Version A and Version B. These versions differed in the following ways:

- Version A: minimalistic, basic font, binary answer choices (Yes/No)
- Version B: colorful (blue/white) interface, playful font, icon/emoji use, Likert scale answer choices (Strongly Disagree → Strongly Agree), examples used to expand on the question posed (i.e., "Do you write things down before acting?")

Overall, by randomly assigning users to one of the two versions and tracking their interactions using Google Analytics, our group aimed to investigate how variations in user interface would influence user engagement.

Key Research Question:
How does the design of a web-based personality quiz affect user engagement and the way users respond to personality-related questions?

Sub-questions:

- Does a more visual and playful format lead to longer interaction times?
- Does a more visual and playful format lead to higher completion rates?
- Do users provide more varied responses when answering via a Likert scale rather than a binary scale?

Experimental Design & Methodology

Our experiment was developed using Shiny in R. We created an interactive web application designed to resemble a personality test, featuring 20 questions in each version. While the questions themselves were identical across both versions, the presentation differed to enable A/B testing. Version A displayed each question in a straightforward format, using standard black and white fonts with no added visual elements. Participants in this version were given binary Yes/No response options. Version B, on the other hand, was designed to be more visually engaging. It included colorful fonts, small images, and additional context beneath each question in the form of explanations or examples. Instead of binary responses, this version provided a range-based (Likert-style) scale for answers. This A/B testing framework was implemented to examine whether the formatting and visual presentation of the test would influence how users interacted with it.

Data Collection

To collect data in a randomized and unbiased manner, we deployed the application under a single link that randomly assigned users to either Version A or Version B upon clicking. The random assignment was implemented using a `determine_version()` function within the Shiny application, which assigned each user to a single version upon launch based on a randomized internal logic, ensuring version allocation was unbiased. This random assignment ensured that each individual had an equal probability of being placed in either group, reducing the risk of selection bias and helping us maintain a representative sample.

Once the application was deployed, each member of our team distributed the link to approximately 15 to 20 individuals. In addition, we shared the link with students enrolled in this course, inviting them to participate by taking as much of the personality test as they wished. Participants were free to stop at any point, and we recorded both partial and complete responses for analysis. For every participant, we tracked several key metrics, including the timestamp, session ID, version assigned (A or B), the number of questions answered, whether the full quiz was completed, and the total time spent on the test. We also recorded whether the participant ultimately submitted their responses. These metrics were chosen as they allowed us to evaluate user engagement and compare behavioral differences between the two versions of the test. This data was linked to an Excel spreadsheet, which

automatically gathered the response data each time a user clicked on the link. This allowed us to compile each user's interaction with the application for the subsequent analysis.

Statistical Analysis & Results

We collected a total of 263 responses, among which:

- Version A had 146 responses, including 92 completed and 54 incomplete, resulting in a completion rate of 63.0137%.
- The average number of questions answered was 15.06, and the average time spent was 77.75 seconds.
- Version B had 117 responses, including 67 completed and 50 incomplete, resulting in a completion rate of 57.2650%.
- The average number of questions answered was 13.54, and the average time spent was 118.09 seconds.

Based on the collected data, we generated the following visualizations.

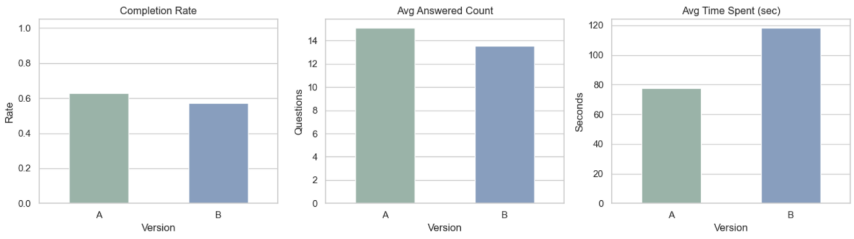

Figure 1: User Performance Comparison Across Versions A and B

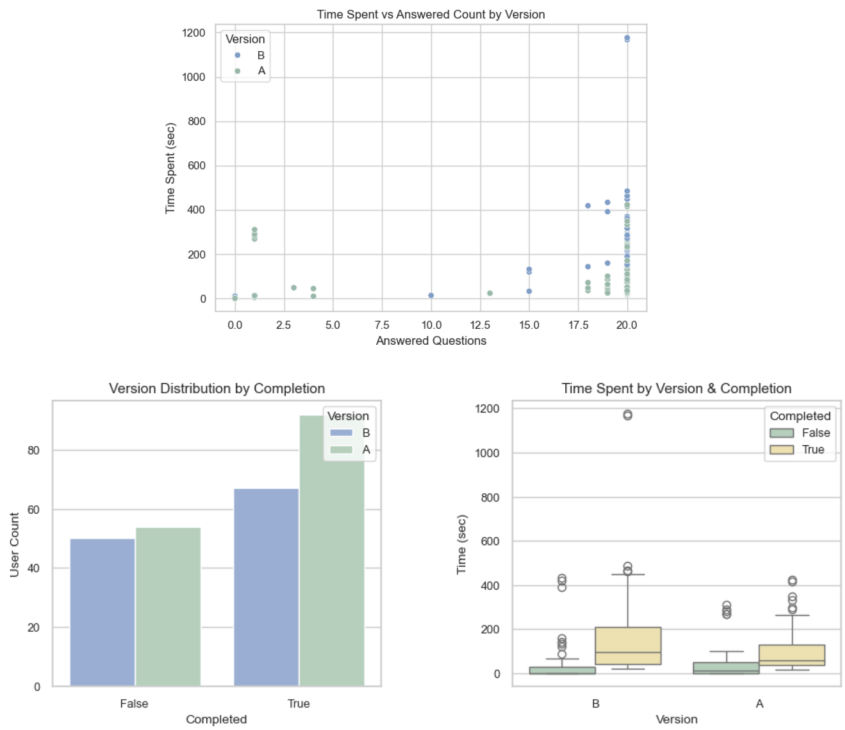

Figure 2: User Completion Behavior and Time Analysis by Version

From the first row of bar charts—Completion Rate, Avg Answered Count, and Avg Time Spent—it can be observed that Group A's completion rate is slightly higher than that of Group B; Group A also answered more questions on average than Group B, indicating that users in Group A tended to answer more questions; Group B's average time spent is significantly longer, exceeding 110 seconds, whereas Group A's average is around 80 seconds.

The second row's scatter plot uses the number of questions answered as the x-axis and time spent as the y-axis, with point color representing different versions. From the result, we can see that there is a high density of points around the 20-question mark in the bottom right, indicating that many participants completed all the questions; Group B users who answered all 20 questions showed a wider spread of time, even exceeding 1000 seconds, suggesting they spent more time thinking or staying on the page; Group A users who answered all 20 questions were mostly concentrated within 100 to 200 seconds, indicating a more efficient and compact completion.

From the visualizations, we can observe differences between Version A and Version B across various indicators. However, considering the differences in participation levels between the two versions, and the need to determine whether these differences are statistically significant, we proceed with a more rigorous analysis using statistical tests.

For the selection of statistical tests, different methods were chosen according to the characteristics of each indicator. The *completion rate* is a categorical variable—whether

the user completed the quiz (`TRUE`/`FALSE`)—corresponding to the two groups (A and B). Therefore, for this kind of association between two categorical variables, a **Chi-Square test** was selected.

For the *answered question count* and *time spent*, the corresponding variables (`answered_count` and `time_spent`) are continuous. Since we aim to compare the means between two independent groups (A vs B), **independent two-sample t-tests** were used to assess whether there are significant differences in means.

At the same time, we are also concerned with the magnitude of the differences, i.e., the *practical significance*, so **Cohen's d** was used to measure the standardized mean difference. The statistical results are shown in the table below.

Test	**Statistic**	**p-value**
Chi-Square Test for Completion Rate	$\text{Chi}^2 = 0.673$	0.4118
T-Test for Answered Questions	t = 1.401	0.1625
T-Test for Time Spent in all users	t = -2.132	0.0345
T-Test for Time Spent in completed users	t = -2.76	0.0071
Cohen's d	d = 0.495	—

From the results, we can see that although Group A had slightly higher completion rates and question counts than Group B, the differences in completion rate and number of questions answered were not statistically significant. However, the p-value for the difference in time spent among all users was 0.0345 (< 0.05), indicating that Group B users spent significantly more time, and that the difference is statistically meaningful.

Furthermore, in a more in-depth analysis focusing only on users who completed the quiz, the p-value for the difference in time spent was 0.0071 (< 0.05), showing that among users who completed the quiz, Group B took significantly longer, and the statistical significance was even stronger.

In the Cohen's d analysis for time spent, the result was $d = 0.495$, which represents a *medium effect size*. This suggests that the difference in time spent between Group B and Group A is not only statistically significant but also practically meaningful.

Interpretation & Conclusion

This study examined how interface design influences user engagement and response behavior in a web-based personality quiz. Two quiz versions were developed: Version A, with a minimalistic layout and binary response options, and Version B, featuring a more visually enriched interface with Likert-scale responses, color elements, and explanatory text. By randomly assigning participants to each version and collecting behavioral metrics, we aimed to assess whether differences in presentation could lead to measurable differences in user interaction.

While initial descriptive statistics suggested that Version A users had higher completion rates and answered more questions on average, further statistical testing revealed that these differences were not significant. However, the time spent on the quiz varied meaningfully between the two groups. Participants assigned to Version B spent significantly more time engaging with the quiz, both among all users ($p = 0.0345$) and specifically among those who completed it ($p = 0.0071$). Notably, the observed medium effect size (Cohen's d = 0.495) suggests that this finding is not only statistically significant but also practically relevant.

These results point to a clear trade-off between engagement depth and completion efficiency. On one hand, Version B's design may have promoted more thoughtful reflection by offering a richer interface and greater response granularity. On the other hand, the increased complexity may have introduced cognitive load, potentially leading to slower progress or even dropout. Therefore, it is essential that the goals of a given survey—whether to encourage deep thinking or maximize response rates—be carefully considered during the design process.

Looking ahead, future studies could enhance this work by incorporating user feedback to better understand perceived usability and engagement. Additionally, examining the quality and consistency of responses across different interface styles would provide further insight into how design choices affect not just participation but also data integrity. Overall, this research underscores the importance of thoughtful, goal-aligned interface design in digital data collection environments.

Challenges & Limitations

Though our experiment effectively showed us that design affects user engagement, a number of limitations must be noted.

First, the two groups were not fully balanced in terms of sample sizes, with more responses from Version A compared to Version B. This potentially brought about slight biases in comparisons even after random assignment.

Secondly, the quiz was taken by participants in uncontrolled settings, which might have impacted focus, duration, and completion. External distraction, variation in devices used (e.g., mobile phones compared to desktops), and variation in internet bandwidth were outside our control and might have introduced noise in data.

Third, even though randomization reduced selection bias, self-selection bias could still exist—those who chose to participate might have been more motivated or interested in personality tests than the broader population.

Also, our participant sample was quite homogenous, mostly consisting of university students, which means that the results cannot be generalized to wider populations.

Most importantly, our measures of engagement were mainly behavioural proxies (com-

pletion rates, duration spent, questions answered) and were not directly measuring participants' cognitive engagement or emotional reaction to varying designs.

Contribution

Task	**Contributors**
Personality Test Creation & Shiny App Deployment	Megha Polavarapu, Jiaqi Guo, Safia
Data Analysis	Yixuan Chen, Jingxi Li
Final Report Writing	All members
Final Report Formatting	Jiaqi Guo

Table 1: *
Summary of individual contributions to different parts of the project.