**1. Completion Funnel Observations**

**Age Group:**

* **Older users drop off more** as the process advances:
  + Start: 32,883 → Step 1: 29,697 → Step 2: 27,848
* **Younger users** follow a similar drop-off trend but **retain more** proportionally.

**Tenure Group:**

* **Long-standing clients** have slightly better retention:
  + Start: 33,300 → Step 1: 30,050 → Step 2: 28,058
* **New clients** drop more steeply:
  + Start: 31,602 → Step 1: 29,155 → Step 2: 28,058

**Insight:** There’s **greater attrition among older and new users**, suggesting a need to tailor UX and content to reduce friction for these groups.

**2. Time Spent per Step & Step Complexity**

**Progression of Time:**

* **Start to Step 1:** ~same time (no significant difference, p = 0.1741)
* **Step 1 to Step 2:** large increase (↑61.6%)
* **Step 2 to Step 3:** small increase (↑7.2%)
* **Step 3 to Confirm:** another big jump (↑32.6%)

**Insight:** **Step 2 and Confirm steps are bottlenecks**, demanding the most user time. Consider UX refinements, better guidance, or breaking them into simpler substeps.

**3. Demographic Differences in Time Spent**

**T-test results:**

* **Age Group:**
  + T = -17.834, p = 0.0000 → statistically significant
  + **Cohen’s d = -0.069** → **very small effect**
* **Tenure Group:**
  + T = -4.737, p = 0.0000 → statistically significant
  + **Cohen’s d = -0.019** → **negligible effect**

**Insight:** While older users spend more time per step and differences are statistically significant (due to large sample sizes), the **practical effect is minimal**.

**4. Backward Navigation Patterns**

**By Step & Variation:**

* Backward rate is **higher in the test group** across all steps except Step 3.
  + E.g., Step 1: Control = 9.6%, Test = 17.9%
* Confirm step has highest overall backward rate in control (37.8%) → drastically reduced in test (11.3%)

**Insight:**

* **Test variant improves navigation clarity** at the final step (confirm).
* But elsewhere, **test version might be causing confusion**, increasing backtracking.

**5. Demographic Influence on Navigation**

**Chi-Square Results (Backward Flag):**

* **Age Group:** χ² = 560.985, p = 0.0000 → significant
* **Tenure Group:** χ² = 13.315, p = 0.0003 → significant

**Effect Sizes (Cramér’s V):**

* **Age Group:** 0.047 → **small**
* **Tenure Group:** 0.007 → **very weak**

**Insight:**

* Age has **some impact** on backward navigation behavior, but **tenure does not**.
* You might consider age-specific guidance or simplifying the flow for older users.

**Overall Key Takeaways**

| **Area** | **Insight** |
| --- | --- |
| **Step Progression** | **Step 2** and **Confirm** require the most time → potential UX bottlenecks. |
| **Demographics** | Older and newer users spend slightly more time and drop off more, but effect sizes are small. |
| **Statistical Tests** | Significant p-values but **small effect sizes** (Cohen’s d, Cramér’s V) → **statistical ≠ practical significance**. |
| **Test vs Control** | Test improves confirm step but may hurt earlier steps (more backward navigation). |

**💡 Recommendations**

1. **Redesign Step 2 and Confirm steps** – simplify content or break into smaller components.
2. **A/B test navigation cues** – especially in earlier steps of the test variant.
3. **Tailor UX for older users** – age has a small but meaningful effect on behavior.
4. **Don’t overreact to p-values** – focus on areas where effect size (impact) is meaningful.
5. **Review copy/UX for early test steps** – higher back navigation suggests confusion.

**Design Effectiveness**

**1. Was the experiment well-structured?**

The experiment appears to be generally well-structured. It included both control and test groups, tracked relevant behavioral metrics (time spent per step, completion funnels, backward navigation), and included segmentation by age and tenure for deeper insights. The use of statistical tests (t-tests, chi-square, effect sizes) indicates a sound analytical approach.

**Supporting Evidence:**

* Clear A/B test structure (e.g., backward navigation rates compared by variation).
* Data tracked across all relevant funnel steps.
* Inclusion of demographic segments (age, tenure).

**Caution:**

* Structure seems valid, but documentation of random assignment or group balancing is not explicitly shown in the dataset. This limits full verification.

**2. Were clients randomly and equally divided between the old and new designs?**

The data suggests an attempt at equal group division (e.g., ~38K in control, ~47K in test at start step), but **no explicit confirmation** of random assignment is provided. The uneven group sizes and possible demographic imbalances raise concerns about whether randomization was strictly applied.

**Supporting Evidence:**

* Uneven start counts (Control: 37,976 vs Test: 47,031) may indicate **imperfect randomization or sample imbalance**.
* Without baseline demographic distributions by group (e.g., % older users in each variant), bias cannot be ruled out.

**Recommendation:**

* Validate or request documentation on **randomization method** and **demographic balance checks** to confirm experiment integrity.

**3. Were there any biases?**

There are **potential biases** related to **age and tenure**, as older and new clients show different behaviors (e.g., time spent, drop-offs, backward navigation). If these demographics were not equally distributed across test and control, they may have influenced outcomes.

**Evidence of bias potential:**

* Significant behavioral differences by age and tenure.
* Cramér’s V and chi-square tests show **age affects backward navigation**, even if the effect size is small.
* Uneven sample sizes between control and test groups.

**Conclusion:**

* **Behavioral biases exist.**
* **Design bias (e.g., group imbalance)** is possible but unconfirmed.

**Duration Assessment**

**Was the timeframe of the experiment (from 3/15/2017 to 6/20/2017) adequate to gather meaningful data and insights?**

Yes, the 3-month duration appears sufficient. The experiment gathered large sample sizes (e.g., 118K+ at start step), enabling robust statistical analysis with high power. The timeframe likely captures a representative range of user behavior.

**Supporting Points:**

* Ample sample size per step (e.g., 86K+ for step\_1).
* Statistically significant results across several tests.
* Time allows for observation of full funnel behavior.

**Recommendation:**

* Ensure no external factors (e.g., seasonality, campaigns) distorted behavior during this timeframe.

**Additional Data Needs**

**What other data, if available, could enhance the analysis?**

Here are **4 key data types** that would meaningfully strengthen the conclusions:

| **Data** | **Why It’s Helpful** |
| --- | --- |
| **Group Assignment Metadata** | Confirms randomization, balance, and avoids confounding variables. |
| **Demographic Distribution by Variant** | Helps test for demographic imbalances between control and test. |
| **Qualitative Feedback (e.g., exit surveys)** | Understands *why* users backtrack or drop off. |
| **Device/Platform Info** | User behavior might vary significantly between mobile vs desktop, which could affect navigation and time spent. |
| **Clickstream Data / Drop-off reasons** | Explains where users struggled, which is key for improving high-time steps like step\_2 and confirm. |

**1. Completion Rate Difference**

**H₀**: Completion rate (Control) = Completion rate (Test)  
**H₁**: Completion rate (Test) > Completion rate (Control)

**Findings:**

* We have **funnel data by step**, but **not the final completion step** for control vs test.
* However, we know:
  + Test group had **higher back navigation** at most steps.
  + No clear evidence of **completion rate increase**.

**Conclusion:**

**Fail to reject H₀.**  
There’s no evidence from the data that **completion rate improved** in the test group — and higher back navigation might indicate **worse** progression.

**2. Completion Rate Threshold (5% Rule)**

**H₀**: Completion rate increase < 5%  
**H₁**: Completion rate increase ≥ 5%

**Findings:**

* Again, **no clear completion rate uplift shown**.
* **Higher drop-off** or backtracking in the test group at most steps.
* No sign that the test version exceeded a 5% gain.

**Conclusion:**

**Fail to reject H₀.**  
You don't have evidence of a **≥5% increase** in completion rate in the test group.

**3. Time Spent per Step (Control vs Test)**

**H₀**: Mean time spent per step is equal between groups  
**H₁**: Mean time spent per step is different between groups

**Findings:**

* There isn’t a direct **time-per-step comparison between control and test**.
* However:
  + A general **increase in time at later steps (step\_2, confirm)** is observed.
  + But **no control/test segmentation** of time makes this inconclusive.

**Conclusion:**

**Inconclusive.**  
You can’t confirm or reject the hypothesis without **mean time per step by control vs test**.

**Recommendation:** Compare mean\_time\_per\_step for control vs test groups directly.

**4. Error Rate (Backward Navigation)**

**H₀**: Error rate (Control) = Error rate (Test)  
**H₁**: Error rate (Control) ≠ Error rate (Test)

**Findings:**

* Backward navigation is **significantly different** between test and control:
  + E.g., Step 1: Test = 17.98%, Control = 9.56%
  + Chi-square test for age and tenure: p < 0.001
* Confirm step is better in test, but all other steps are worse.

**Conclusion:**

**Reject H₀.**  
There is a **significant difference** in backward navigation (error rate) between test and control groups.

**5. Step Comparison (Step 1 vs Step 2 Time, within sessions)**

**H₀**: Mean time Step 1 = Mean time Step 2  
**H₁**: Mean time Step 1 ≠ Mean time Step 2

**Findings:**

* **Step 1 mean time**: 54.02
* **Step 2 mean time**: 87.29
* T-test between steps: **T = 1.36, p = 0.1741**
* No statistically significant difference (p > 0.05)

**Conclusion:**

**Fail to reject H₀.**  
Even though the **mean time increases**, the **difference is not statistically significant** based on the t-test provided.

**Summary Table**

| **Hypothesis** | **Result** | **Conclusion** |
| --- | --- | --- |
| 1. Completion Rate Difference | Fail to reject H₀ | No evidence test outperforms control |
| 2. 5% Completion Rate Threshold | Fail to reject H₀ | No evidence of ≥5% uplift |
| 3. Time Spent (Control vs Test) | Inconclusive | Missing time-per-step by group |
| 4. Backward Navigation Rate | Reject H₀ | Test and control differ significantly |
| 5. Step 1 vs Step 2 Time | Fail to reject H₀ | No significant difference |