DATA ANALYTICS REPORT

Mobile Site Survey A/B Test Analysis

Hypothesis Testing | Statistical Analysis | Actionable Insight

Jordan Foltz
Jordy3338@gmail.com
linkedin.com/in/jordan-foltz

March 2026

Introduction

A software company tested a new mobile-first help website against the original version.

Users who visited the site on a mobile device were randomly assigned to one of two groups. The control group saw the original website. The test group saw the new mobile-first redesign. After browsing, each user rated their satisfaction on a 1 to 5 scale, where 1 is least satisfied and 5 is most satisfied. The experiment ran from July 1 to July 23, 2025.

The developers wanted to detect any meaningful difference in satisfaction scores before committing to further revisions. The dataset contains 3,365 responses across both iOS and Android devices.

Hypotheses

The test is two-tailed. The developers want to know whether the new site produces any change in satisfaction — positive or negative — not just improvement. This is the appropriate framing when the direction of the effect is unknown.

> **NULL HYPOTHESIS (H_0)**
> The new mobile-first website produces no difference in user satisfaction ratings compared to the original site. Any observed difference between the two groups is the result of random variation.
>
> $H_0 : \mu_{test} = \mu_{control}$

> **ALTERNATIVE HYPOTHESIS (HA)**
> The new mobile-first website produces a difference in user satisfaction ratings compared to the original site. The difference is real and not the product of random variation.
>
> $Ha : \mu_{test} \neq \mu_{control}$

Why two-tailed? A two-tailed test is used because the developers want to detect any significant change in either direction. If the new site performed significantly worse, that finding would be equally important to act on.

Data Summary

The experiment collected 3,365 survey responses across both groups during July 2025.
The control group contains 1,666 users and the test group contains 1,699 users. The groups are closely matched in size, which supports a fair comparison. Approximately 69.8% of respondents used iOS and 30.2% used Android.

Metric	Control (Original)	Test (New Site)
Users (n)	1,666	1,699
Mean rating	**3.2767**	**3.3373**
Variance	1.4603	1.5063
Std deviation	1.2084	1.2273
Median rating	**3.0**	**3.0**

Rating Distribution

Looking at how ratings are distributed across the 1 to 5 scale reveals nuance that summary statistics alone do not capture.

Rating	Control	Test	Difference
1 (worst)	11.3%	9.7%	-1.6pp
2	14.2%	15.9%	+1.7pp
3	24.5%	24.4%	-0.1pp
4	35.2%	30.9%	-4.3pp
5 (best)	**14.7%**	**19.1%**	**+4.4pp**

WHAT THE DISTRIBUTION TELLS US

The test group shows more 5-star ratings (19.1% vs 14.7%) but also more 2-star ratings (15.9% vs 14.2%) and fewer 4-star ratings (30.9% vs 35.2%). The shift is mixed rather than consistently positive. Both groups share a median of 3.0, confirming the overall experience did not change meaningfully.

Two-Sample Welch T-Test

I used a two-sample Welch t-test assuming unequal variances to compare mean satisfaction scores.

This test is appropriate because ratings are continuous, the two groups are independent, and the variances differ between groups (1.46 for control vs 1.51 for test). The Welch adjustment accounts for this difference and produces a more accurate result than a standard t-test assuming equal variances.

The test compares the mean ratings of both groups and determines whether the observed difference of 0.06 points is large enough to be statistically meaningful, or whether it could plausibly occur by random chance.

Calculation Results

Statistic	Value	Notes
Control mean	3.2767	Original site
Test mean	3.3373	New mobile site
Mean difference	+0.0605	Test minus control
Control variance	1.4603	
Test variance	1.5063	
Observations (control)	1,666	
Observations (test)	1,699	
Degrees of freedom	3,363	Welch approximation
t-statistic	-1.4419	
p-value (two-tailed)	**0.1494**	Threshold: 0.05
Reject H_0?	**No**	**0.1494 > 0.05**

Statistical Conclusion

The p-value of 0.1494 is well above the 5% significance threshold. The null hypothesis is not rejected.

There is insufficient statistical evidence to conclude that the new mobile site produces a different level of user satisfaction compared to the original. The observed mean difference of 0.06 points on a 5-point scale is not large enough to rule out random variation as the explanation.

> **INTERPRETING THE P-VALUE**
>
> A p-value of 0.1494 means that if the two sites were truly identical in user experience, we would expect to see a mean difference this large roughly 15 times out of 100 experiments. That is not rare enough to conclude the effect is real. We need a result this extreme to occur fewer than 5 times out of 100 to reject the null. Additionally, Cohen's d of 0.05 indicates a negligible effect size, confirming the redesign had minimal practical impact on user experience.

Actionable Insights

The new mobile site did not produce a statistically significant change in user satisfaction.

This is a fail-to-reject result, not a failure. It means the data does not provide enough evidence to conclude the redesign improved the user experience. The developers should treat this as a signal to revisit the design before committing to a full rollout.

0.1494	+0.06	0.05
P-Value **Threshold: 0.05**	Mean Difference **Out of 5-point scale**	Cohen's d **Negligible effect size**

Recommendation

Do not deploy the new mobile site in its current form.

The redesign did not demonstrate a statistically significant improvement in user satisfaction. The p-value of 0.1494 means the observed difference is consistent with random variation. Deploying the new site without evidence of improvement carries the risk of disrupting an experience that users currently rate at 3.3 out of 5.

The rating distribution data does offer one constructive signal. The test group produced more 5-star ratings (19.1% vs 14.7%), which suggests certain users responded positively to specific features of the redesign. Further research into what drove those high ratings could inform a more targeted revision.

FINAL VERDICT

Fail to reject Ho. The new mobile site does not produce a statistically significant difference in user satisfaction at the 5% significance level. Further design iteration is recommended before reconsidering deployment.

Tools and Methods

Two-sample Welch t-test	Unequal variances assumed
Pivot table (Microsoft Excel)	A/B Test Analysis ToolPak
Cohen's d (effect size)	Two-tailed hypothesis testing
Rating distribution analysis	Descriptive statistics