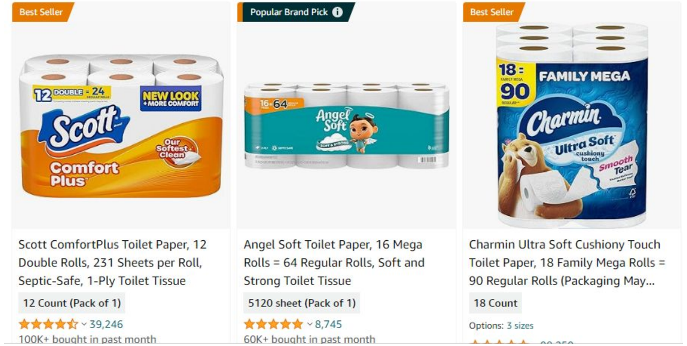
Optimizing Amazon’s E-commerce Recommendations: Impact of Grouped Displays

Team 2
Yue Sun, Kaia Hu, Jiuyuan Xie

Background



- Amazon recommends certain items with product labels.
  - Best Seller
  - Top Pick
  - Amazon Pick
  - Overall Pick
- Those recommended products are arranged randomly and scattered throughout the results shown.
- This arrangement design can be confusing and time consuming when users browse through the search results.

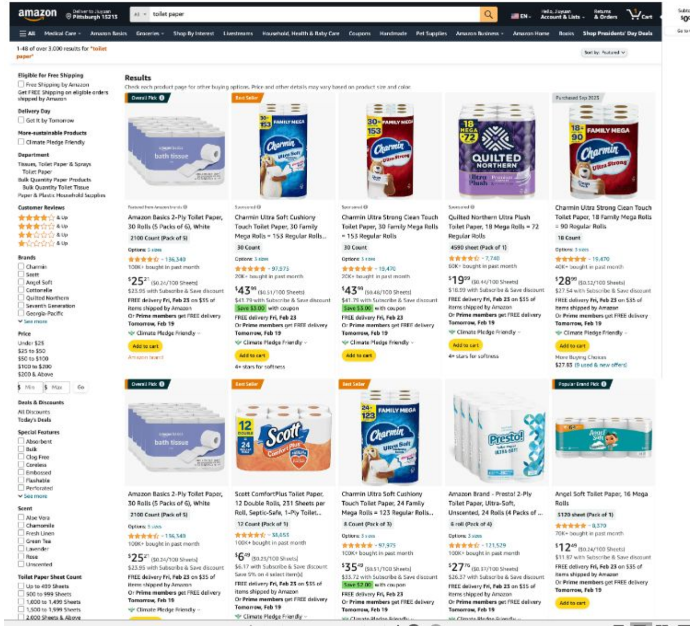
Research Question

Does grouping and prioritizing the picked items together at the top of the search results more likely to lead higher selection rates of the recommended products?

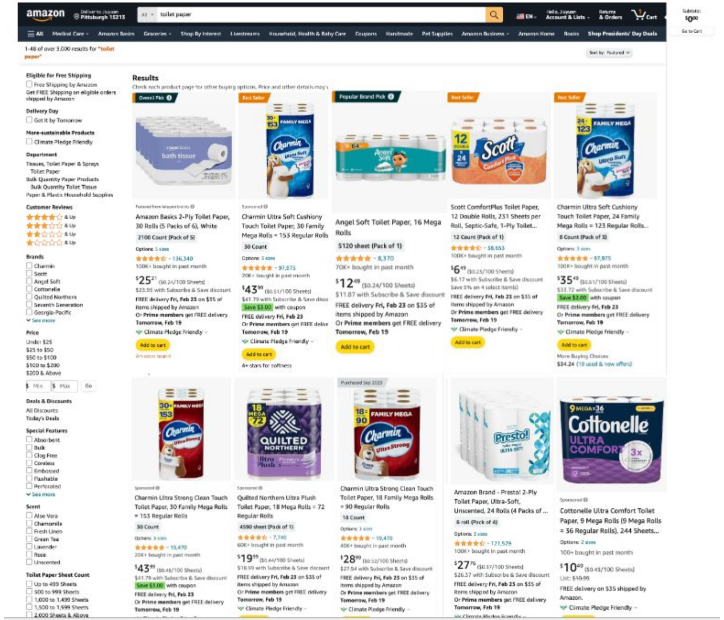
Hypothesis

Grouping and prioritizing labeled products at the top of search results will lead to higher selection rates of recommended products, compared to a control group exposed to the standard, scattered arrangement of labeled products in search results.

Treatment Implementation - Survey Design

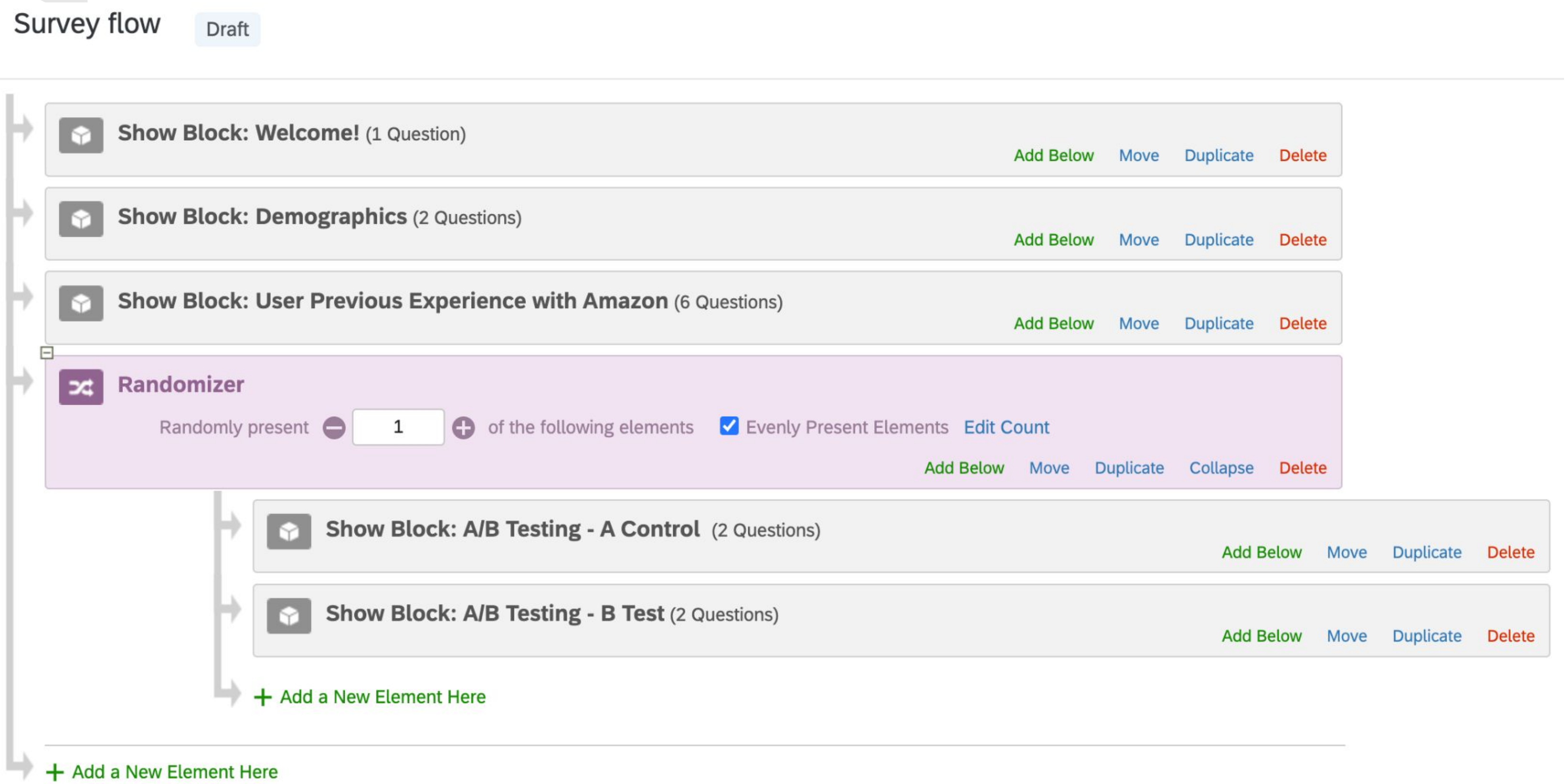


Displayed Product Page in **Control** Group

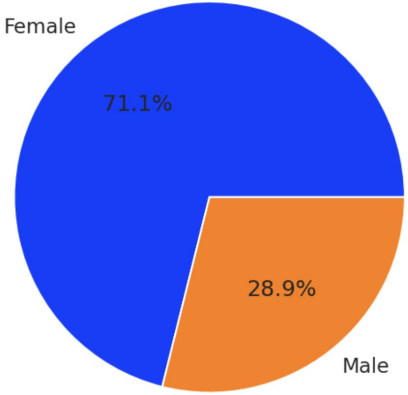


Displayed Product Page in **Treatment** Group

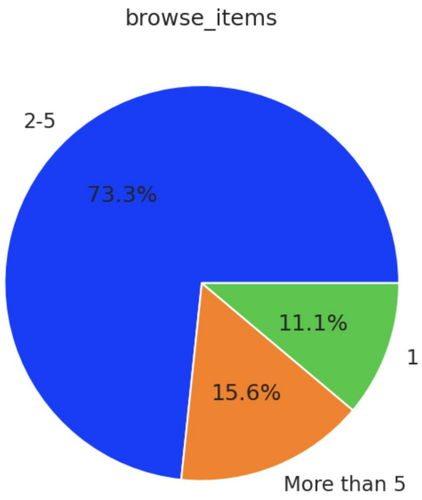
Treatment Implementation - Randomization



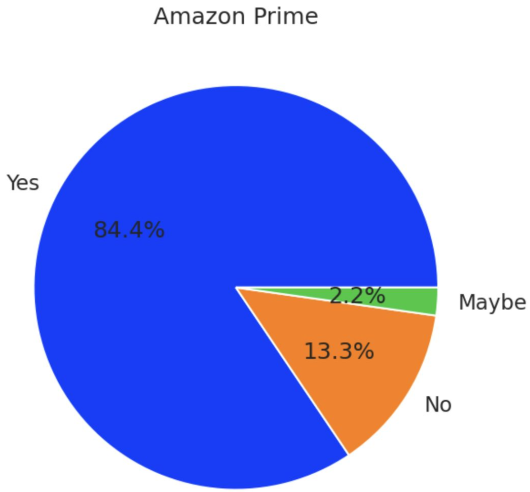
EDA - Demographics & User Experience on Amazon



Gender

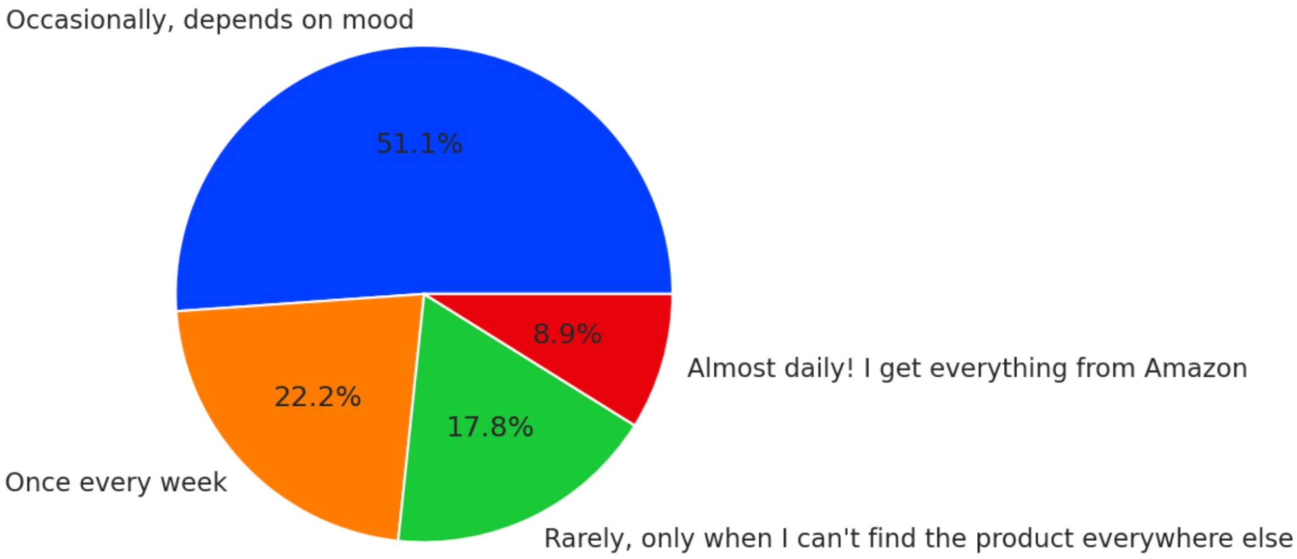


User Browsing Habit



Amazon Prime Usage

EDA - Demographics & User Experience on Amazon



User Shopping Habit

Results - T tests

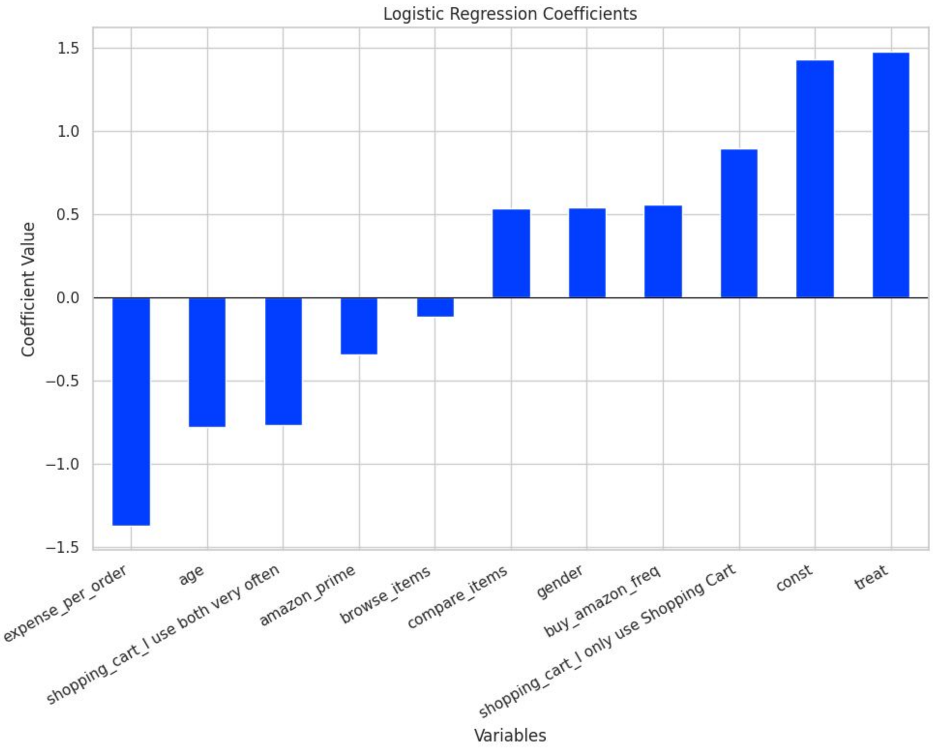
Treatment Effect	T Stats	P Values
0.2035	1.3827	0.1738

We ran a most basic T test on the results, without any feature engineering, to understand our data. The P value is bigger than the significance level of 0.05, which indicates that we should not reject the null. However, we decided to run more models to look into it further.

Results - Logistic Regression

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.4341	2.932	0.489	0.625	-4.312	7.181
age	-0.7805	0.680	-1.148	0.251	-2.113	0.552
amazon_prime	-0.3460	1.169	-0.296	0.767	-2.637	1.945
browse_items	-0.1140	0.192	-0.593	0.553	-0.491	0.263
buy_amazon_freq	0.5592	0.414	1.350	0.177	-0.253	1.371
compare_items	0.5326	0.553	0.963	0.335	-0.551	1.616
expense_per_order	-1.3734	0.788	-1.743	0.081	-2.918	0.171
gender	0.5427	1.066	0.509	0.611	-1.547	2.632
shopping_cart_I only use Shopping Cart	0.8937	1.785	0.501	0.617	-2.604	4.391
shopping_cart_I use both very often	-0.7707	1.875	-0.411	0.681	-4.446	2.905
treat	1.4787	0.807	1.833	$\boxed{0.067}$	-0.102	3.060

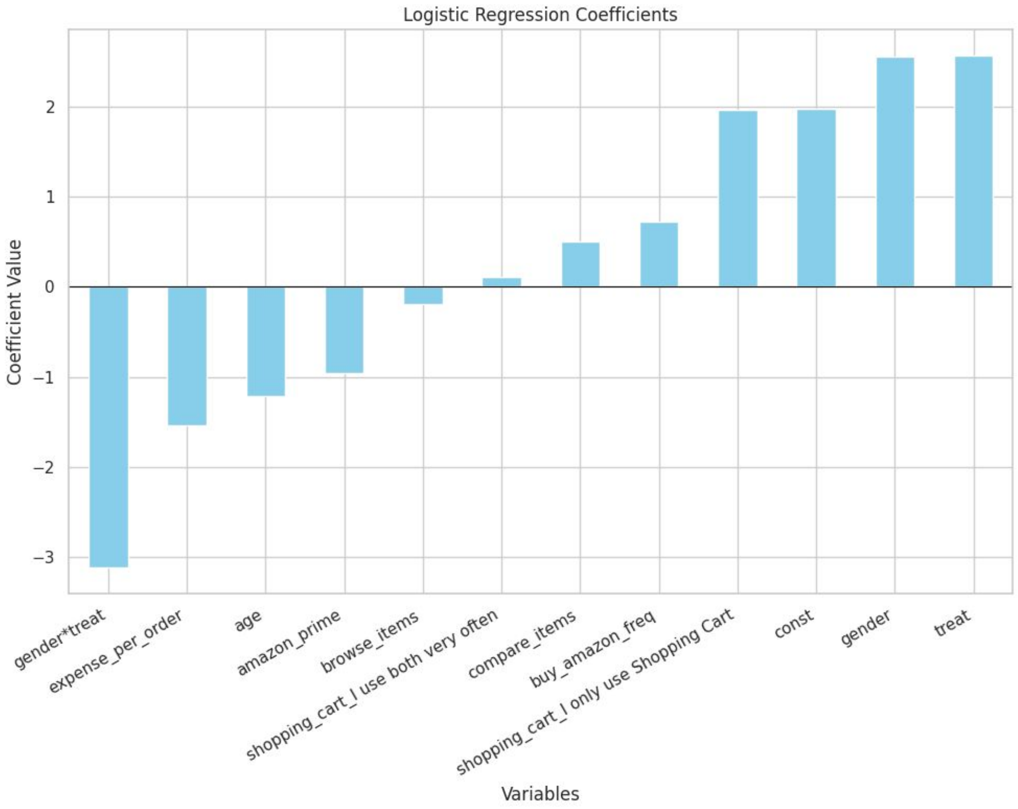
The P value has improved, but still remained insignificant; let's see if adding an interaction term help.



Results - Interaction Term with Logistic Regression

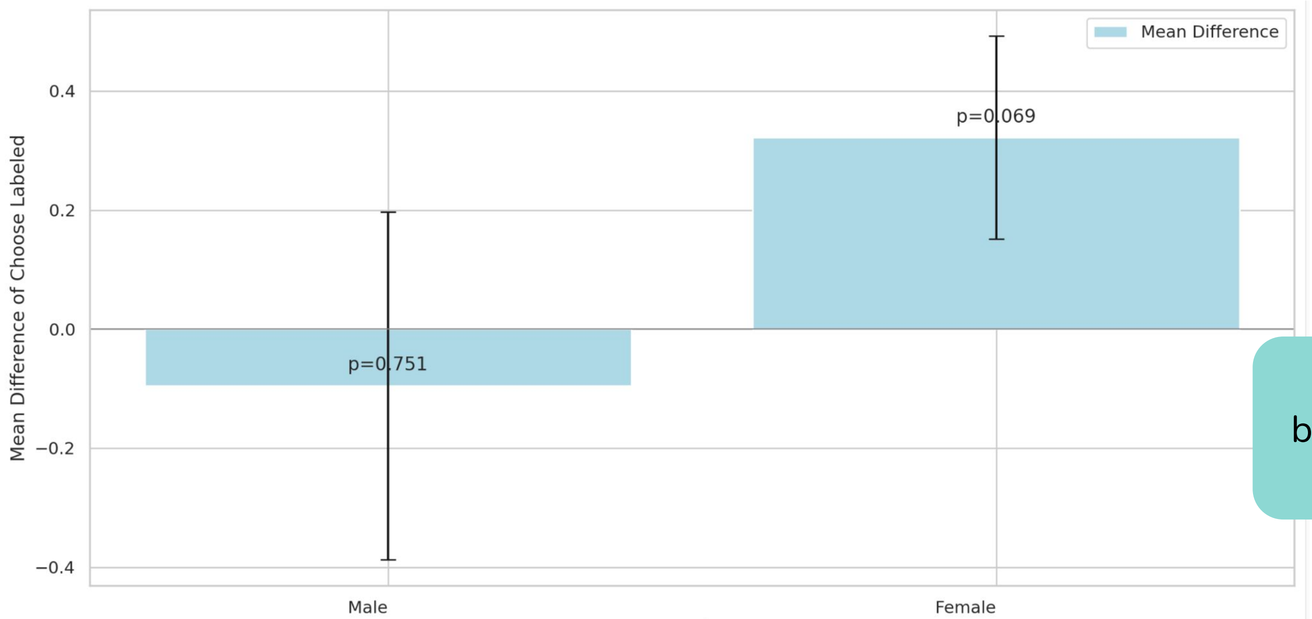
	coef	std err	z	P>|z|	[0.025	0.975]
const	1.9770	3.132	0.631	0.528	-4.162	8.116
age	-1.2098	0.772	-1.568	0.117	-2.722	0.302
amazon_prime	-0.9595	1.285	-0.747	0.455	-3.477	1.558
browse_items	-0.1896	0.213	-0.889	0.374	-0.608	0.229
buy_amazon_freq	0.7270	0.480	1.515	0.130	-0.214	1.668
compare_items	0.5041	0.611	0.825	0.409	-0.693	1.701
expense_per_order	-1.5360	0.871	-1.764	0.078	-3.242	0.170
gender	2.5536	1.812	1.409	0.159	-0.999	6.106
gender*treat	-3.1132	2.025	-1.537	0.124	-7.083	0.856
shopping_cart_I only use Shopping Cart	1.9614	1.997	0.982	0.326	-1.954	5.876
shopping_cart_I use both very often	0.1047	2.072	0.051	0.960	-3.956	4.166
treat	2.5666	1.145	2.242	0.025	0.323	4.810

After adding the gender-treatment interaction term, we can observe that the p-value drops to under significance level.



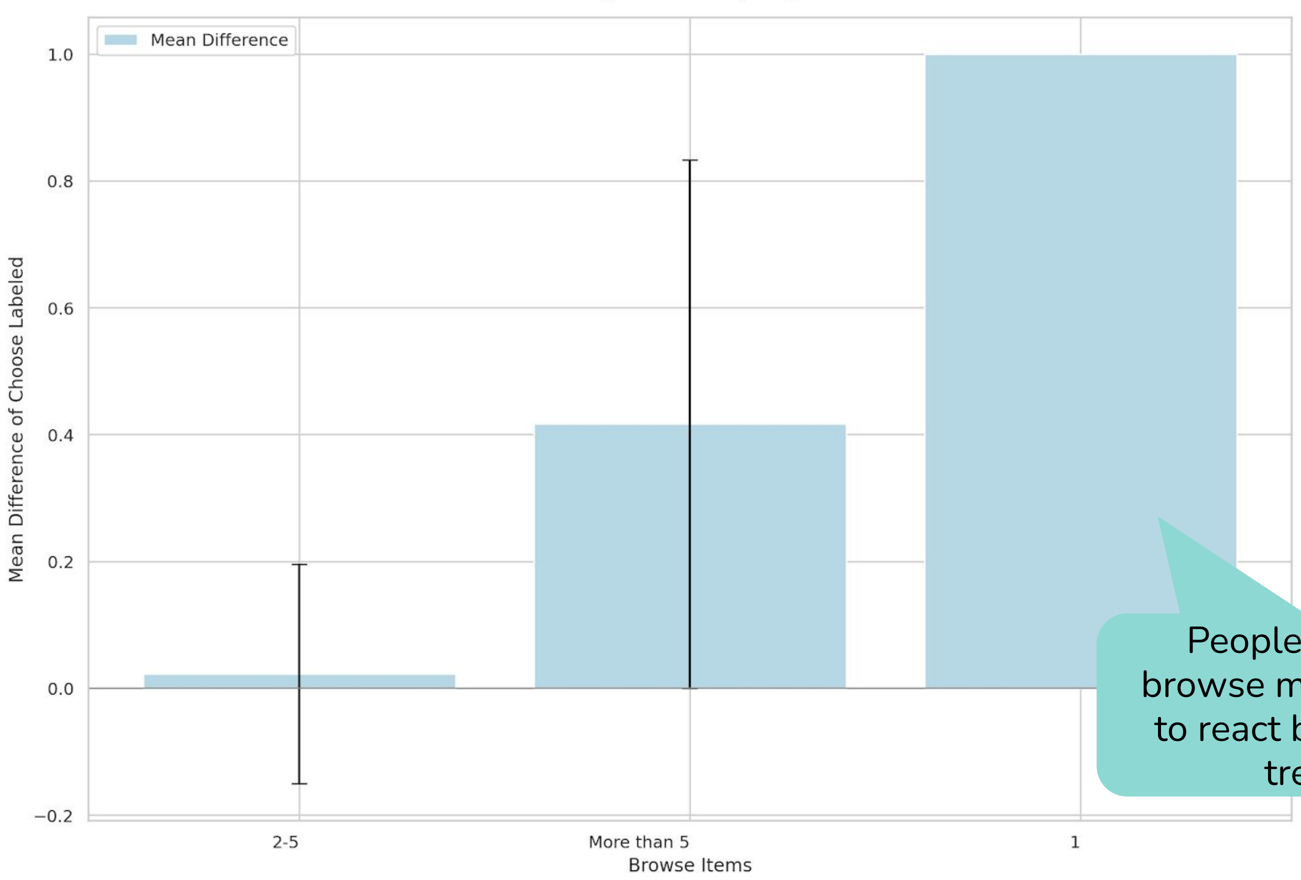
T tests with Segments-Gender

	control_samples	treat_samples	total_samples	Effect	T-test	P-value
gender-['Male']	6	7	13	-0.095	-0.326	0.751
gender-['Female']	17	15	32	0.322	1.885	0.069



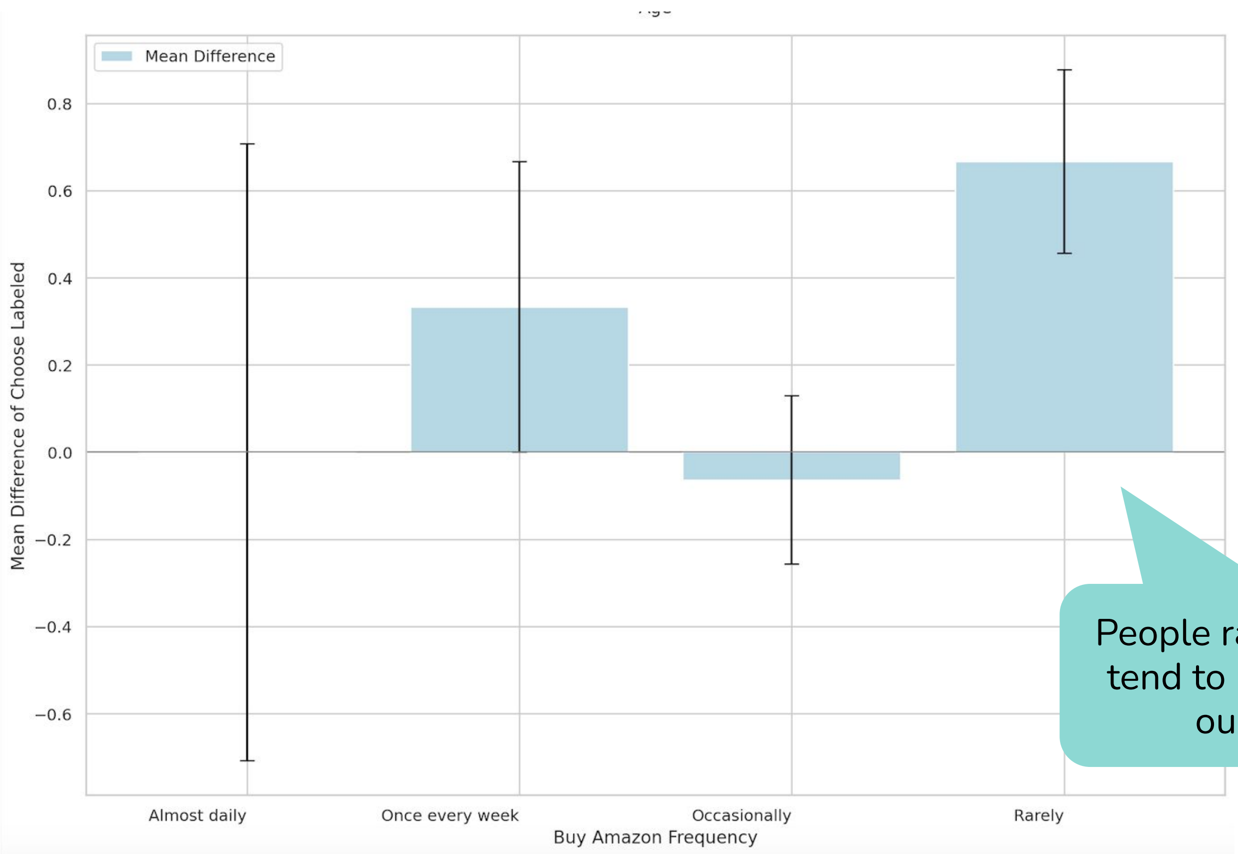
Female tend to react better with our treatment than male

T tests with Segments- #Browse items



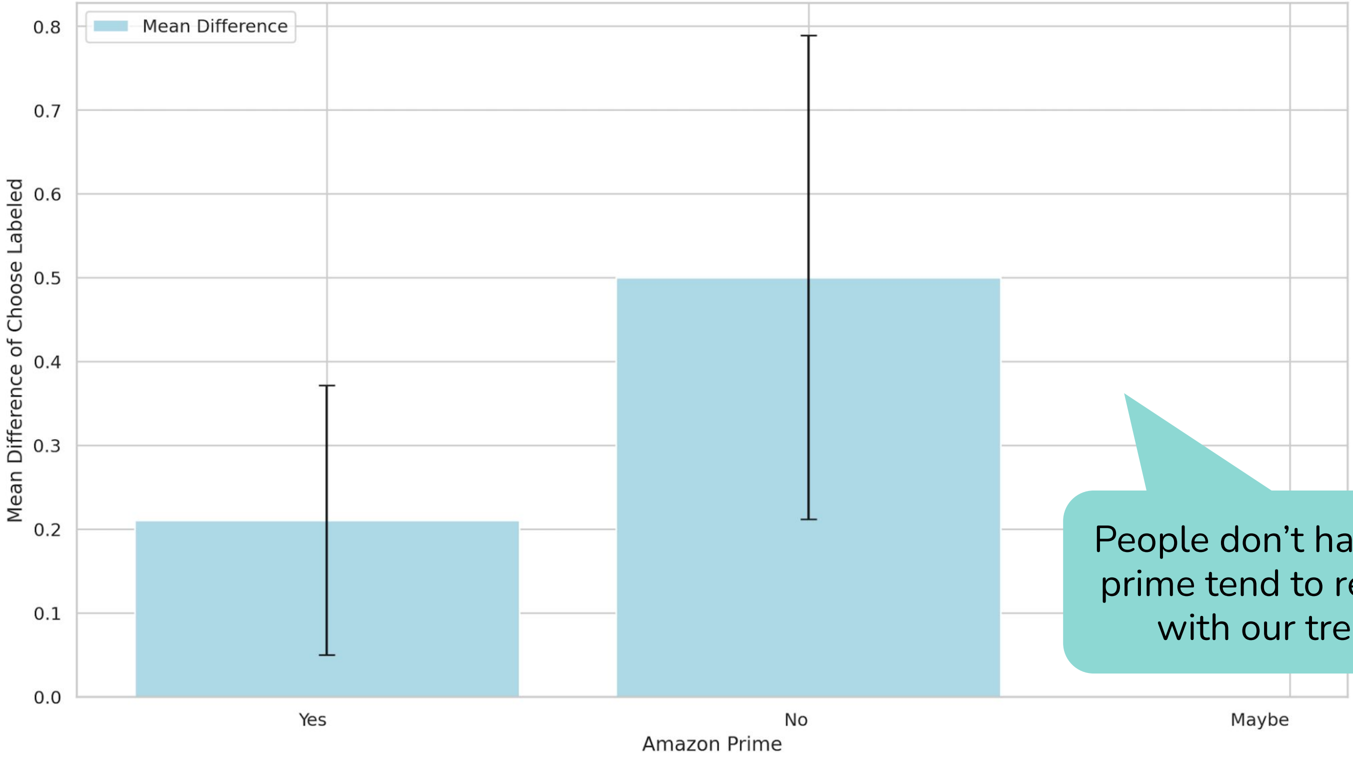
People don't like to browse many items tend to react better with our treatment

T tests with Segments- Use amazon Freq



People rarely use amazon tend to react better with our treatment

T tests with Segments- Amazon Prime



People don’t have amazon prime tend to react better with our treatment

Evaluation - Sample size calculation

Desired Increase	20%
Observed CTR	26/43 =0 .6
Variance	0.6*(1-0.6)
Sensitivity	0.72-0.6

16 * (0.6)*(1-0.6)/(0.12)**2
=261.5

Results Interpretation

- Model without interaction term: treatment has a p-value of 0.067 (> 0.05).
  - While there may be an association between the treatment and the outcome (choosing labeled items), the evidence is not strong enough to assert this with high confidence
- Model with interaction term (gender): treatment has a p-value of 0.025 (< 0.05).
  - When controlling for the interaction between gender and treatment, the treatment has a significant effect on the outcome

Some Other Interesting Findings

- The treatment appears to have a more positive effect on **female**, with a near-significant p-value, suggesting a possible trend that females might be more influenced by the treatment than males. However, the difference is not statistically significant.
- **Participants who browse fewer items (1 item)** show a higher effect size compared to those who browse more items. However, the difference is not statistically significant.
- Users who **rarely buy** on Amazon show the largest effect size suggesting they may be more influenced by the treatment. However, the difference is not statistically significant.
- **Non-prime members** showed a marginally higher effect from the treatment compared to the prime members, the difference is not statistically significant.

Conclusion

- The data does not provide enough evidence to reject the null hypothesis in favor of the alternative across the entire sample.
- While there are indications that certain segments of the population may be positively influenced by the treatment of grouping and prioritizing labeled products, these effects do not consistently reach statistical significance.
- Further research with large sample sizes or different methodologies may be necessary to explore these trends more deeply.