Report: DA106 – A/B Testing for GloBox

1. Context

Client

GloBox is a reasonably fast-growing online marketplace that sells unique and high-quality products from around the world. Its mainstay boutique fashion and high-end décor categories generate 85% of revenues. The nascent food and drink category rounds up remaining 15%. Revenues grow at +11% p.a. but are slowing.

Problem statement

GloBox needs to reinvigorate its growth story ahead of its initial public offering (IPO) on a US exchange in spring next year. To justify its anticipated valuation, it needs to show momentum in growing its revenues.

Envisioned solution and execution

The food and drink category is growing rapidly (c.+30% p.a.) while the mainstay categories are maturing (c.+5% p.a.). GloBox believes growing its food and drink category further is necessary to reaccelerate overall growth.

GloBox plans to implement a program that runs a **new banner that highlights its key food and drink products** (placed at the top of its mobile homepage for maximum visibility) to shift user behavior over time. To start, it will run a randomized controlled test (RCT) as an A/B test, on the new banner.

The A/B test is a key part of this strategy. If the test shows that the new banner leads to sustainably higher revenues, GloBox will make significant investments in this area and even expand its plethora of growth initiatives for its food and drink category. GloBox will then be able to articulate the food and drink category as a viable growth engine, and reaffirm investor confidence in the company's future for the IPO.

Alternative growth strategy

Armed with a high margin (albeit cyclical) mainstay categories and a fortress-like balance sheet with ample cash and zero net debt, GloBox can pursue acquisitions of rival platforms and other vertical subjects (e-wallets, logistics) in select geographies.

However, acquisitions can muddy its growth story prior to the IPO. Absence of exchange-tradeable stock as an acquisition currency can be suboptimal from a shareholder optimization perspective, according to its financial advisors and working capital providers.

Desired impact

GloBox's pivot towards food and drink can spur much-needed growth in its business, and to increase revenues, market share, and customer stickiness. This would help to smooth revenue variability and increase its valuation prior to the IPO.

Stakeholder analyses

Below outlines the scope and requirements of key stakeholders involved in the new banner program, setting the tone for the A/B test.

Role	Domain	Priorities
Data Analyst	Data and the analysis.	Collect and analyze data to understand user behavior. Run hypotheses testing. Communicate and make recommendations.
Product manager	High level insights. KPIs (metrics).	Set goals for projects. Metric-manage user conversion rates and $/user sales. Align all stakeholders on priorities and goals. Execute resource allocation and prioritization.
UX designer	Detail UX design.	Syncretize user research with the couture of digital experiences. Pitch and execute designs.
Head of marketing	Marketing and research.	Discover valuable niches for targeting. Benchmark metrics against the market by user segments. Communicate marketing plans.

Author

- Alan Kong, Data Analytics candidate January 2023 intake, Masterschool.

2. Planning and setup of A/B test

Objectives of A/B test

Obtain statistically-backed evidence from test subjects and answer the following:

- Will the new banner confidently and significantly boost proxies for revenue growth, namely, conversion rate (CR), spend per user ($/user) and conversion economics ($/conversion)?
- Any learning effects such as change aversion and novelty effect?
- What are the feasible iterations on the new banner post A/B test?

Timeframe

- 13 days, from January 25th to February 6th, inclusive. Typically quiet period for retail. Possibly greater likelihood to capture unbiased test subjects.

Conversion

- This occurs when a test subject makes one (or more) purchases. A converted test subject must be unique, and may make one or more purchases.

Metric choice (owner: product manager)

Propensity measure:

- (Add-to-cart) conversion rate (CR) = total conversions / total test subjects.

Volume measure:

- Spend per user ($/user) = Total spend in $ / total test subjects.
- Conversion economics ($/conversion) = Total spend in $ / total conversions.

Rates measures such as click-through-rate (CTR) not measured.

Field methodology

- Mobile website users only.
- **Onboarding:** randomizer tool captures and assigns mobile users (as test subjects) to control (A) or treatment (B) groups. This will run daily from 10:00am and 10:00pm (at user local times). **Control group A** will be the **baseline**.
- Point at which a user becomes a test subject is the user's join date.
- Only those in **treatment group B** (often referred to as **test** or **test group**), will see the **new banner** (that highlights food and drink items). Baseline users will see the default banner featuring mainstay products (similar to other users at large).
- Conversion is indifferent to product categories, may even consist of exclusively mainstay products.

Data collection

- Purchasing activity data and personal data (device type, gender and country) are collected for user and marketing research by UX and Marketing.
- Regulatory restrictions on user data privacy and other technical limitations mean some test subjects may lack the complete set of personal data.

Effect size and related practical significance and substance test
A less technical test called the substance test is done alongside hypotheses testing.

- Relative change (difference in values between test and baseline as proportion of baseline value) as choice effect size that is comparable across the metrics.
- Substance test is carried to see if effect size of the new banner exceeds a threshold (pass) or falls short (fail).
- Management sets 5% minimum threshold for effect size on overall platform. A higher 10% on sub-segments, on higher perceived investment and execution risk.
- Given baseline growth of c.+30% p.a. in the food and drink category, management feels, an effect size of 30% or more is considered large for their practical significance, 10-30% (medium) and 5-10% (small). Below 5% is insignificant or poor if the change is negative.
- A more niche program may show small practical significance but still fails, if for example, it achieves 7% effect size but fails the higher 10% (or higher) threshold.
- Co-verify the substance test with hypotheses test.

Cohen’s d had been looked at as an alternative practical significance metric (difference in values as multiple of standard error) but not be used due to:

- Lack of precedence.
- Difficulty in convincing an active-management family office (and board member) to use the metric. The family office does not view risk from a variability stand point, rather the risk of permanent capital loss.

Significance level and power of the test, tails

Power analysis:

- Given **relatively low marginal cost** of running A/B test, the strengths of the test are set at **0.80 or 80% statistical power** and **0.05 or 5% significance level**.
- On that basis, a power analysis is run to determine if the sample size is sufficient for the test to at least detect the observed effect size.
- If the minimum required sample size (to detect observed effect size) is at or less than actual sample size, then the actual sample is deemed sufficient as its effective Minimum Detectable Effect or MDE is lower versus the observed effect size (on relative change basis).
- Thresholds set by the management will serve as Floor MDE, to place a ceiling on sampling and prevent excessive pursuit in A/B test. This aims to limit unforeseen adverse effects to user behavior on the platform due to A/B test.
- Floor MDE and threshold (for effect size) are referred to interchangeably.

Two-tailed z-test and t-test will be run as appropriate and on the basis of the significance level.

3. Execution risks, tradeoffs and test propriety

Weaknesses (identified risks)	Tradeoffs
• Limiting A/B test to mobile users may artificially depress conversion rates. • Non-subscriber test subjects whose subsequent activity (on desktop) may go unrecorded.	• Limiting to mobile users may reduce the risk of bias related to higher premeditated purchasing intent often on desktop. • A/B test on both mobile and desktop users may generate more data for more robust test but may amplify the bias.
• Lower sensitivity to the banner as A/B test is run during weak seasonality.	• Risk of business disruption from A/B test is lower during periods of lower traffic.
• Lower sensitivity from just testing one banner due to generalized tastes differences across user segments.	• Gaps that manifest can be cost-effective steer for iteration, instead of deploying resources to 'get it right' the first time.
• Click-through rates (CTR) not tested, and situation not helped by the lack of collection of product-level data. • Spurious insights on perceived statistical significance could be led coincidently by mainstay products, not basket diversity. Unable to assess cannibalization of mainstay products by food and drink items.	• Marketing and UX agree to waive gaps temporarily for speedy execution. • Lack of basket level product data due to data being stored separately by a newly onboarded third-party software vendor. • Collection of such product data will go live for future tests, as new API is in production.
• Regulatory risk of data collection such as infringement of EU's GDPR.	• Systematic cure via use of automatic user-consent waivers. • Small number of test subjects lack some personal data, but no immaterial impact for Marketing and UX research purposes.

Test propriety

A/B test is fit and proper for this exercise as:

- There is only one variable, being the new banner to distinguish the test group from baseline and is not complex.
- It is important to be aware of the potential risks associated with the test so that the results can be interpreted accordingly.

4. Recommendation guidance

Weaknesses and tradeoffs of the A/B test cast a “fog” that needs to be accounted for with sensible and relatively strict recommendation options.

Launch the experience

- if confident that the new banner has significant positive influence on all metrics.
- Statistical significance and substance test are mutually consistent.

Perform minor iteration and then launch the experience

- if confident the new banner has significant positive influence on CR only. But $/user and $/conversion can improve with minor tweaks and selective targeting, without having to increase sampling.

Continue iterating with design of the experience and run a new A/B test

- if confident the new banner has significant positive influence on CR only. But $/user and $/conversion can improve with material iterations with or without having to increase sampling.

Abort and not launch the experience or the program

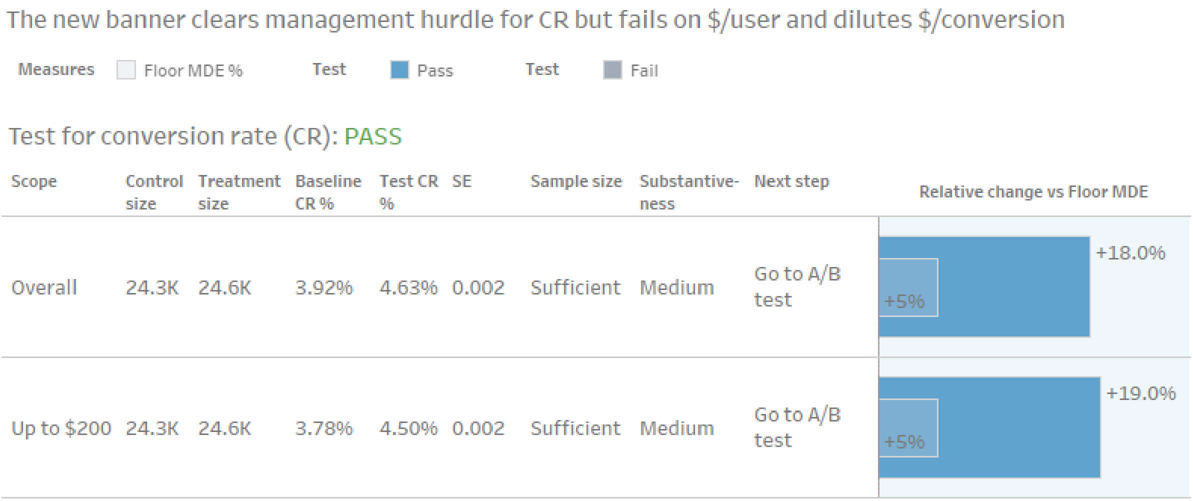
- if not confident the new banner has significant positive influence on any of the metrics (especially CR). Few viable options to iterate.

Understandably the middle two options are nuanced rejections of the new banner, while keeping the option open to build a new design to boost overall revenues.

5. Preview analysis: substance test

Overview and substance test

- New banner seems to raise CR by +18%. This is the only metric that passes the substance test and beats management threshold or Floor MDE (+5%).



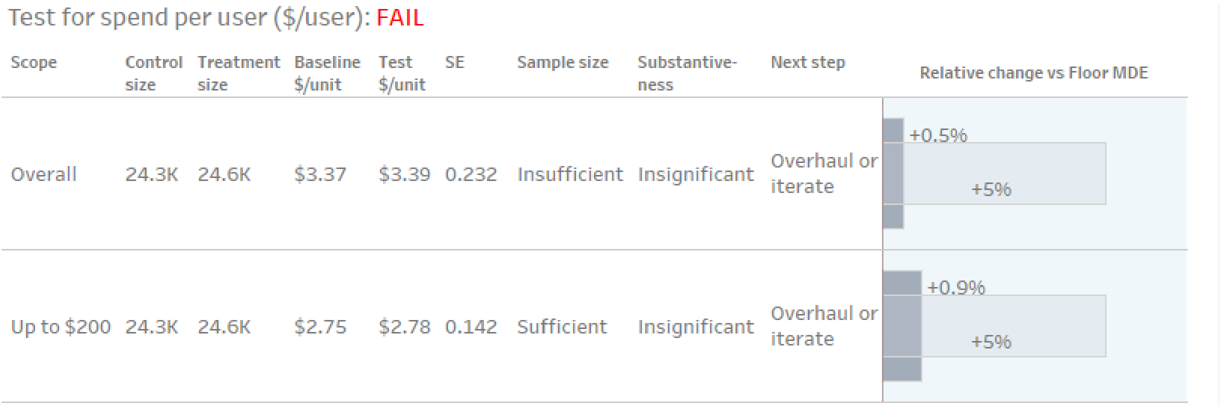
The new banner clears management hurdle for CR but fails on $/user and dilutes $/conversion

Measures □ Floor MDE %   Test 🟦 Pass   Test ▨ Fail

Test for conversion rate (CR): PASS

Scope	Control size	Treatment size	Baseline CR %	Test CR %	SE	Sample size	Substantive-ness	Next step	Relative change vs Floor MDE
Overall	24.3K	24.6K	3.92%	4.63%	0.002	Sufficient	Medium	Go to A/B test	+18.0% +5%
Up to $200	24.3K	24.6K	3.78%	4.50%	0.002	Sufficient	Medium	Go to A/B test	+19.0% +5%

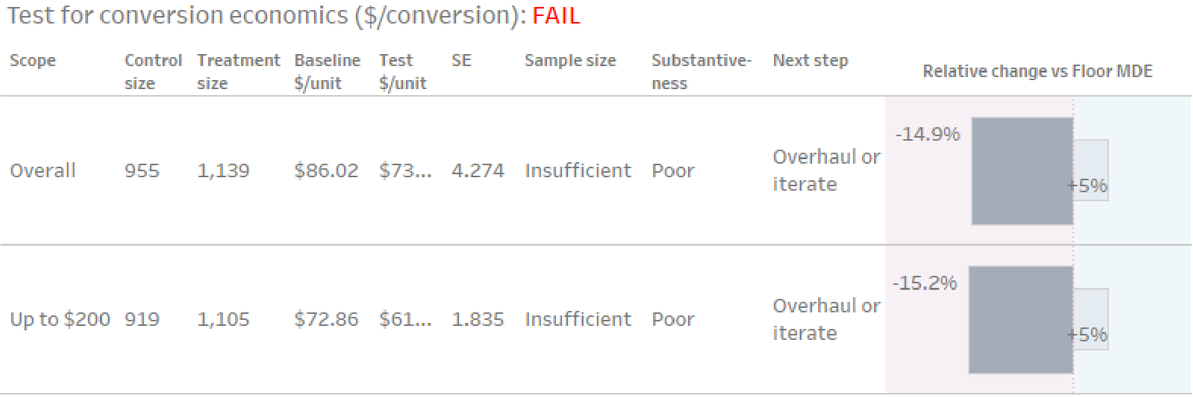
- New banner is muted for $/user.



Test for spend per user ($/user): FAIL

Scope	Control size	Treatment size	Baseline $/unit	Test $/unit	SE	Sample size	Substantive-ness	Next step	Relative change vs Floor MDE
Overall	24.3K	24.6K	$3.37	$3.39	0.232	Insufficient	Insignificant	Overhaul or iterate	+0.5% +5%
Up to $200	24.3K	24.6K	$2.75	$2.78	0.142	Sufficient	Insignificant	Overhaul or iterate	+0.9% +5%

- Sharp -15% dilution in $/conversion in the test group offsets effects of higher CR.



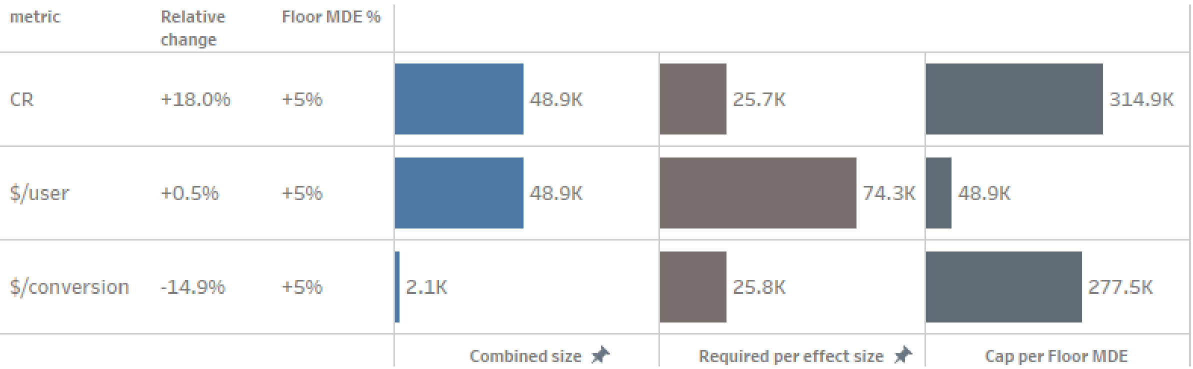
Test for conversion economics ($/conversion): FAIL

Scope	Control size	Treatment size	Baseline $/unit	Test $/unit	SE	Sample size	Substantive-ness	Next step	Relative change vs Floor MDE
Overall	955	1,139	$86.02	$73...	4.274	Insufficient	Poor	Overhaul or iterate	-14.9% +5%
Up to $200	919	1,105	$72.86	$61...	1.835	Insufficient	Poor	Overhaul or iterate	-15.2% +5%

Power analysis and sample sizes

- 48,943 of mobile users/visitors to the website are onboarded as test subjects. Effectively 50:50 split between the baseline (24,343) and test group (24,600).
- Per the prior-defined significance level (0.05) and statistical power (0.80) the test is sufficiently sensitive to detect effect size in CR but insufficient for other metrics. Below compares sample sizes versus metric-related requirements.

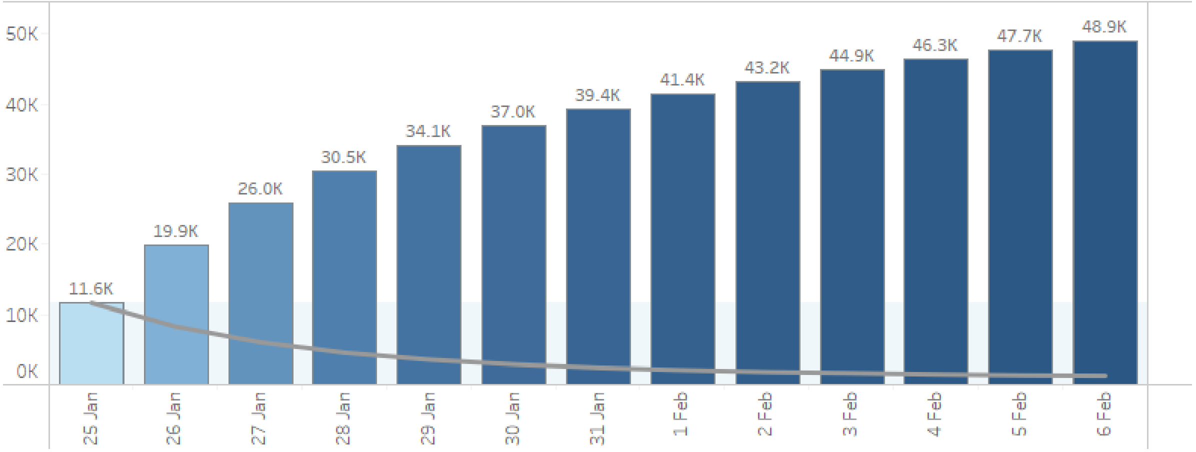
Test with sample size sufficiently sensitive to detect CR. Insufficient when downstream to $/conversion



metric	Relative change	Floor MDE %	Combined size	Required per effect size	Cap per Floor MDE
CR	+18.0%	+5%	48.9K	25.7K	314.9K
$/user	+0.5%	+5%	48.9K	74.3K	48.9K
$/conversion	-14.9%	+5%	2.1K	25.8K	277.5K

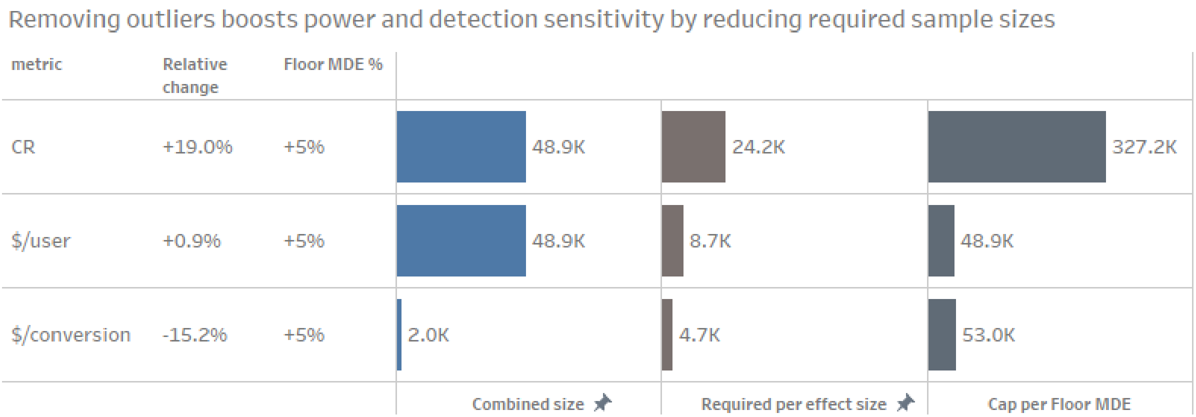
- The observed onboarding ramp below shows that it takes three days to reach the required 26K cumulative test subjects, for the test to detect effect size +18% for CR.

Sample reaches sufficiency in 3 days to detect changes in CR. Need to ramp up more for $/conversion



- The grey line on the chart reads number of test subjects added daily over the timeframe shown. The ramp starts at a rapid pace, and slows gradually towards the sample size obtained by the end of the timeframe.
- Ramp speeds and lengths of timeframe can be customized to reach desired sample sizes if possible.

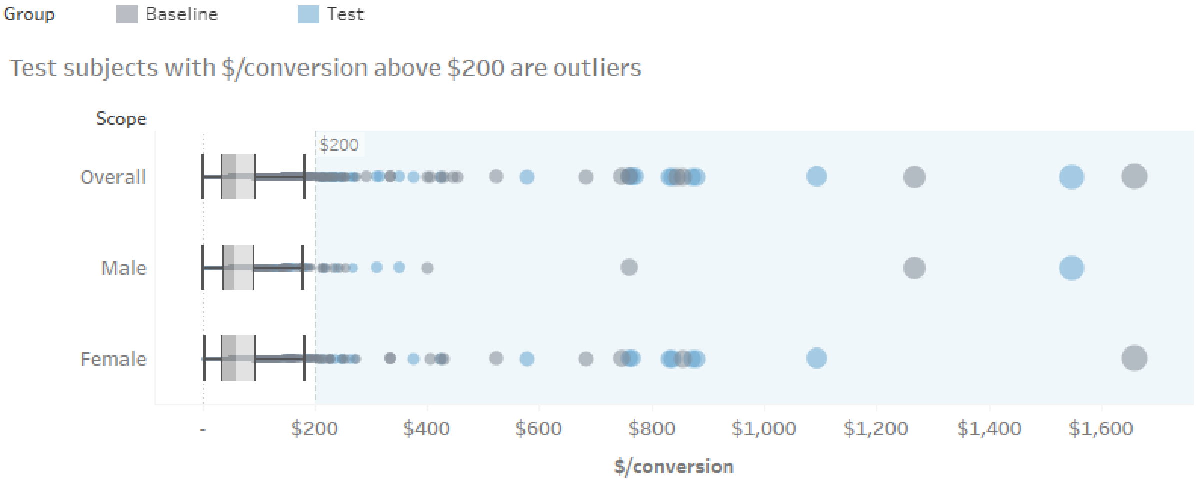
- Per below by removing only 70 outliers (i.e. conversion with spend above $200 each) can cure the test's sensitivity for $/user. Sample size requirement drops (from combined 74K) to 8.7K, far below the 48.9K sample size.



Removing outliers boosts power and detection sensitivity by reducing required sample sizes

metric	Relative change	Floor MDE %	Combined size	Required per effect size	Cap per Floor MDE
CR	+19.0%	+5%	48.9K	24.2K	327.2K
$/user	+0.9%	+5%	48.9K	8.7K	48.9K
$/conversion	-15.2%	+5%	2.0K	4.7K	53.0K

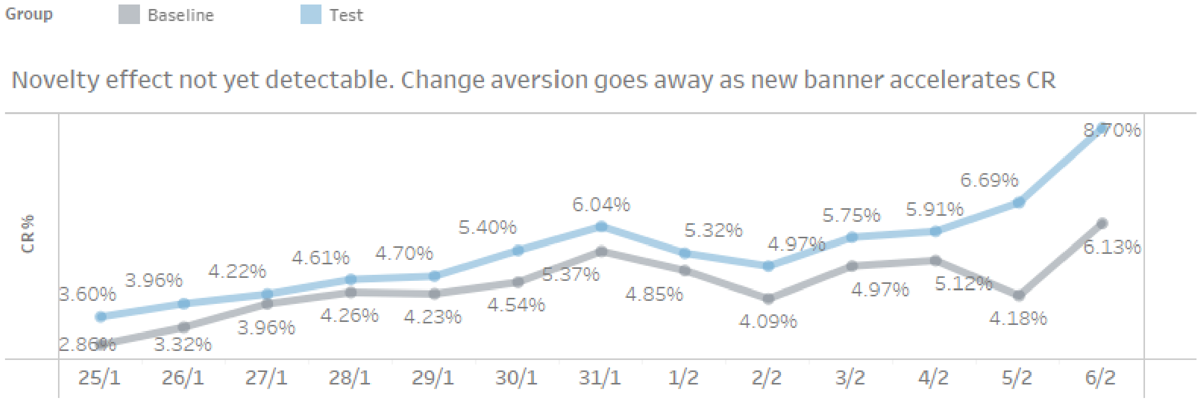
- More importantly is the $/conversion (from which $/user is derived). Here, the sample size is still insufficient despite the exclusion of outliers.
- Boxplots below illustrate the outliers. Choice of $200 cut-off aligns consistently in various sub-segments such as gender, shown here.



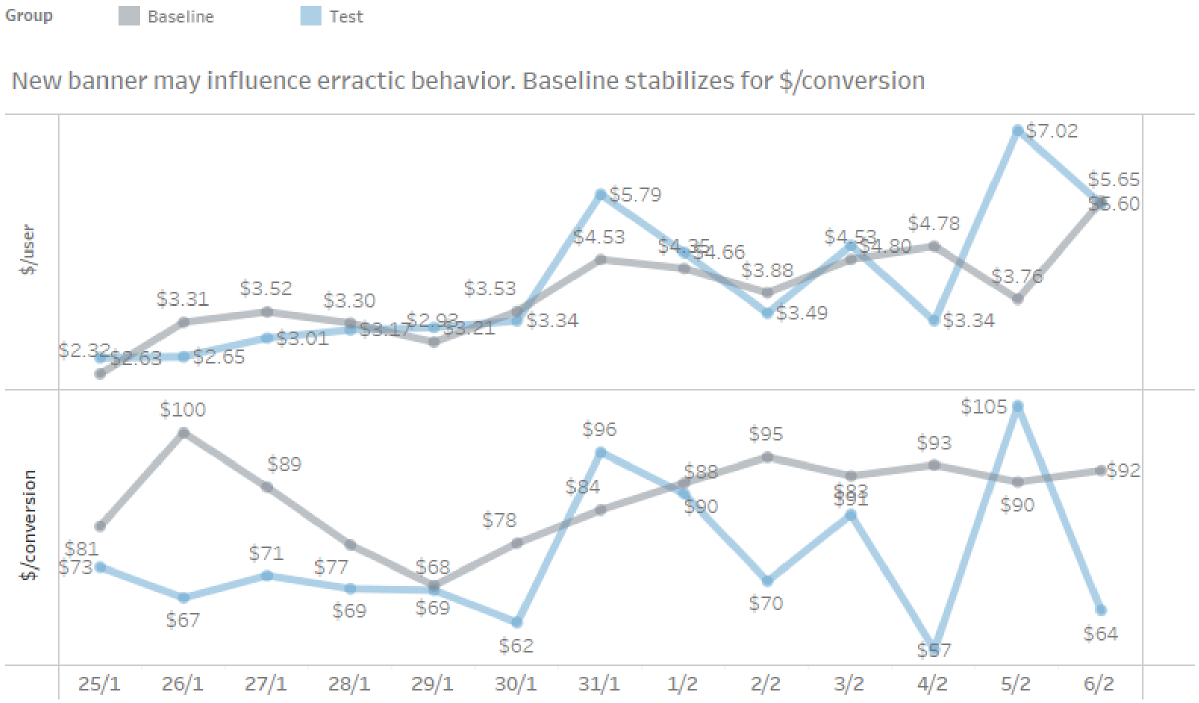
Test subjects with $/conversion above $200 are outliers

Learning effects from the lens of metric evolution

- Change aversion disappears as the new banner accelerates CR in the final days.
- Timeframe may be too short. Risk of missed opportunity to adequately study stabilization, peak metric or novelty effects in CR.



- Much noise in test group for $/conversion and its derivative, $/user.
- Baseline stabilization in $/conversion (to $90+ area) is not quite unexpected.



Initial conclusion and critique

- **Little practical evidence to support launch** of the program (new banner).
- To continue with the program, some **iteration or overhaul is needed** to **boost $/conversion** and solve for upstream sampling for CR to **obtain large enough sample size** downstream for $/conversion.
- $/user, as a derivative of $/conversion, requires less attention.
- **Caution** when inferring from product-agnostic $/conversion metric, lacking context on product-diversity or (pre-checkout) order of item selections.
- **Unaccounted noise (but may not be)** in $/conversion in the test group could be the result of interactions between various effects such as inter-category cannibalization effect, complementary goods effect and/or changes in purchasing intent (on exposure to the banner). Insufficient information to infer anything on this currently.

6. A/B Test results

Overview

The default position or the null hypothesis (H0) presupposes no evidence that the new banner leads to any significant change in CR, $/user or $/conversion in the test group versus the baseline.

The alternative hypothesis (H1) contradicts H0 and suggests there is evidence that the new banner is effective on those metrics in the test group.

Both full sample ('Overall') and the subset without outliers ('Up to $200') are tested.

CR

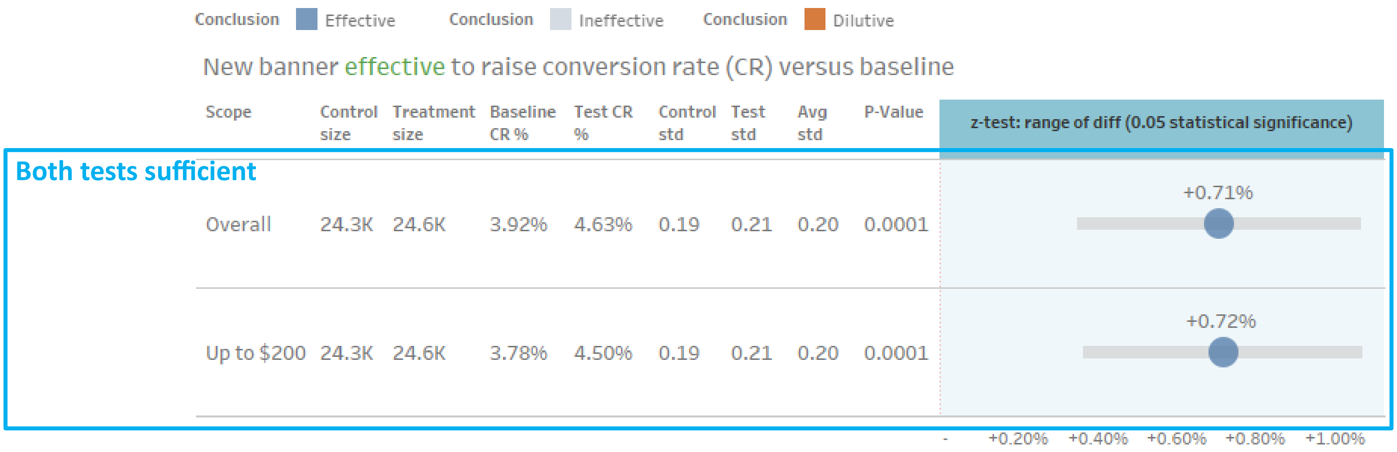
- Both cases are sufficient tests (power = 0.80, significance level = 0.05).
- Metric type = proportion %. Test type = 2-tail z-test, unequal variance.

Hypothesis

- H0: Diff (= CR_{test} - $CR_{baseline}$) = 0 i.e. Diff (if any) observed, due to chance.
- H1: Diff (= CR_{test} - $CR_{baseline}$) ≠ 0 (not zero) i.e. DIff due to behavior change.

Conclusion:

- Diff in CR estimated to be in interval below per 0.95 (1- 0.05) confidence level.
- Probability H0 is true given Diff observed and interval is **p-value 0.0001**, which is very unlikely and is **materially below 0.05 significance level**.
- Thus, **reject H0**. It is unlikely the Diff is due to chance.
- Strong evidence, the new banner changes user behavior to raise CR. Substance test reaffirmed.



Conclusion ■ Effective Conclusion ■ Ineffective Conclusion ■ Dilutive

New banner effective to raise conversion rate (CR) versus baseline

Both tests sufficient

Scope	Control size	Treatment size	Baseline CR %	Test CR %	Control std	Test std	Avg std	P-Value	z-test: range of diff (0.05 statistical significance)
Overall	24.3K	24.6K	3.92%	4.63%	0.19	0.21	0.20	0.0001	+0.71%
Up to $200	24.3K	24.6K	3.78%	4.50%	0.19	0.21	0.20	0.0001	+0.72%

$/user

- Only 'Up to $200' case is sufficient test (power = 0.80, significance level = 0.05).

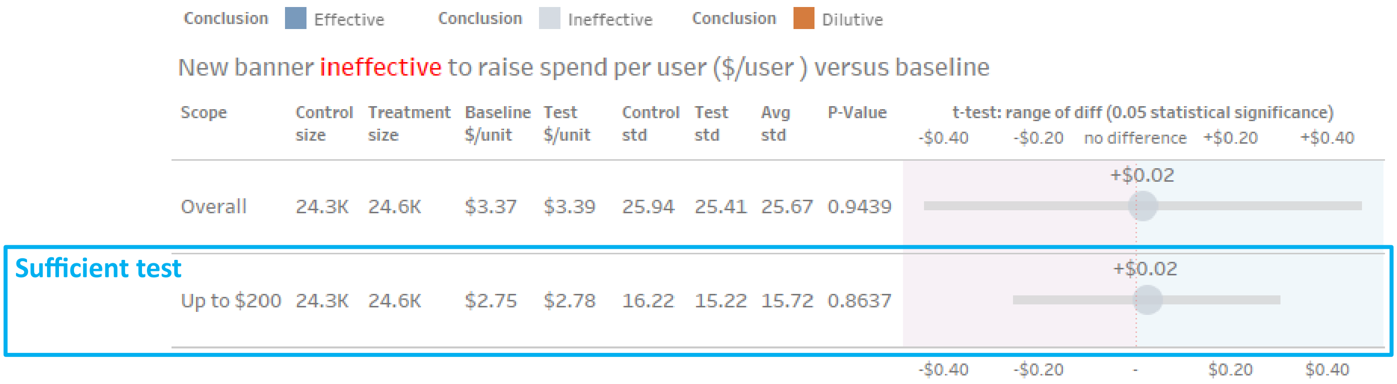
- Metric type = means $. Test type = 2-tail t-test, 48,605 degrees of freedom, unequal variance.

Hypothesis

- H0: Diff $(= \$/\text{user}_{\text{test}} - \$/\text{user}_{\text{baseline}}) = 0$ i.e. Diff (if any) observed, due to chance
- H1: Diff $(= \$/\text{user}_{\text{test}} - \$/\text{user}_{\text{baseline}}) \neq 0$ (not zero) i.e. DIff due to behavior change

Conclusion:

- Diff in $/user estimated to be in interval below per 0.95 (1- 0.05) confidence level.
- Probability H0 is true given Diff observed and interval is **p-value 0.8637**, which is very likely and is **materially above 0.05 significance level**.
- Thus, **fail to reject H0**. It is likely the Diff is due to chance.
- The new banner is insignificant on $/user. Substance test reaffirmed.



Conclusion 🟦 Effective   Conclusion ⬜ Ineffective   Conclusion 🟧 Dilutive

New banner ineffective to raise spend per user ($/user) versus baseline

Scope	Control size	Treatment size	Baseline $/unit	Test $/unit	Control std	Test std	Avg std	P-Value	t-test: range of diff (0.05 statistical significance)
Overall	24.3K	24.6K	$3.37	$3.39	25.94	25.41	25.67	0.9439	+$0.02
Sufficient test Up to $200	24.3K	24.6K	$2.75	$2.78	16.22	15.22	15.72	0.8637	+$0.02

$/conversion

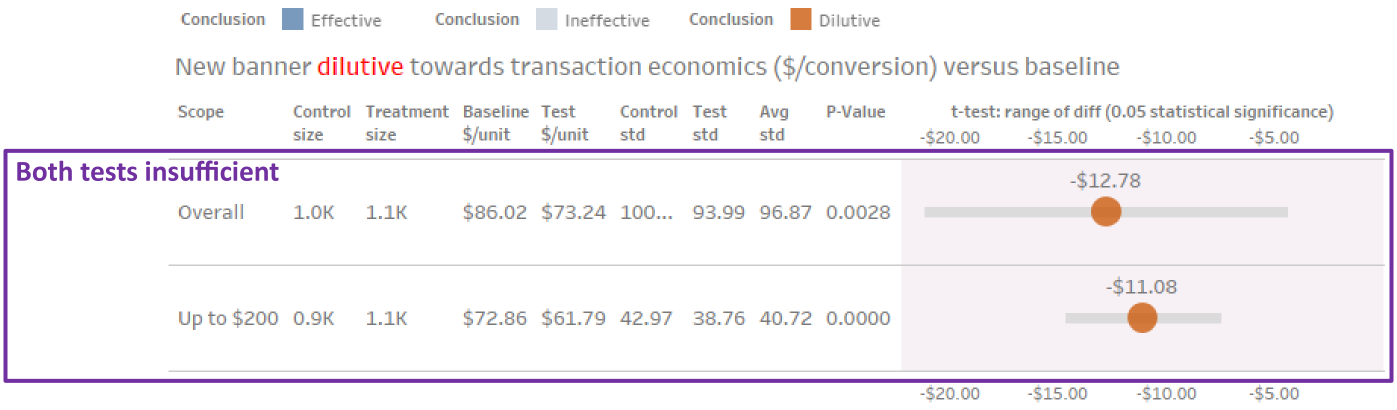
- Both cases are insufficient tests. Conclusion is only tentative.
- Metric type = means $. Test type = 2-tail t-test, 1,978 degrees of freedom, unequal variance.

Hypothesis

- H0: Diff (= $\$/conv_{test}$ - $\$/conv_{baseline}$) = 0 i.e. Diff (if any) observed, due to chance
- H1: Diff (= $\$/conv_{test}$ - $\$/conv_{baseline}$) ≠ 0 (not zero) i.e. DIff due to behavior change

Conclusion:

- Diff in $/conv estimated to be in interval below per 0.95 (1- 0.05) confidence level.
- Probability H0 is true given Diff observed and the interval is **p-value close to 0** , which is very unlikely and is **materially below 0.05 significance level**.
- Tentatively, **reject H0**. It is unlikely the Diff is due to chance.
- Tentative evidence, the new banner changes user behavior poorly and dilutes $/conversion. Substance test reaffirmed tentatively.



Conclusion ■ Effective Conclusion □ Ineffective Conclusion ■ Dilutive

New banner dilutive towards transaction economics ($/conversion) versus baseline

Both tests insufficient

Scope	Control size	Treatment size	Baseline $/unit	Test $/unit	Control std	Test std	Avg std	P-Value	t-test: range of diff (0.05 statistical significance)
Overall	1.0K	1.1K	$86.02	$73.24	100...	93.99	96.87	0.0028	-$12.78
Up to $200	0.9K	1.1K	$72.86	$61.79	42.97	38.76	40.72	0.0000	-$11.08

7. Initial impression

On the whole evidence suggests that the **new banner** program has insignificant influence on $/user. Any campaign aimed at driving traffic to the website (to be converted) is as good as doing nothing. The program as it is, is not value accretive.

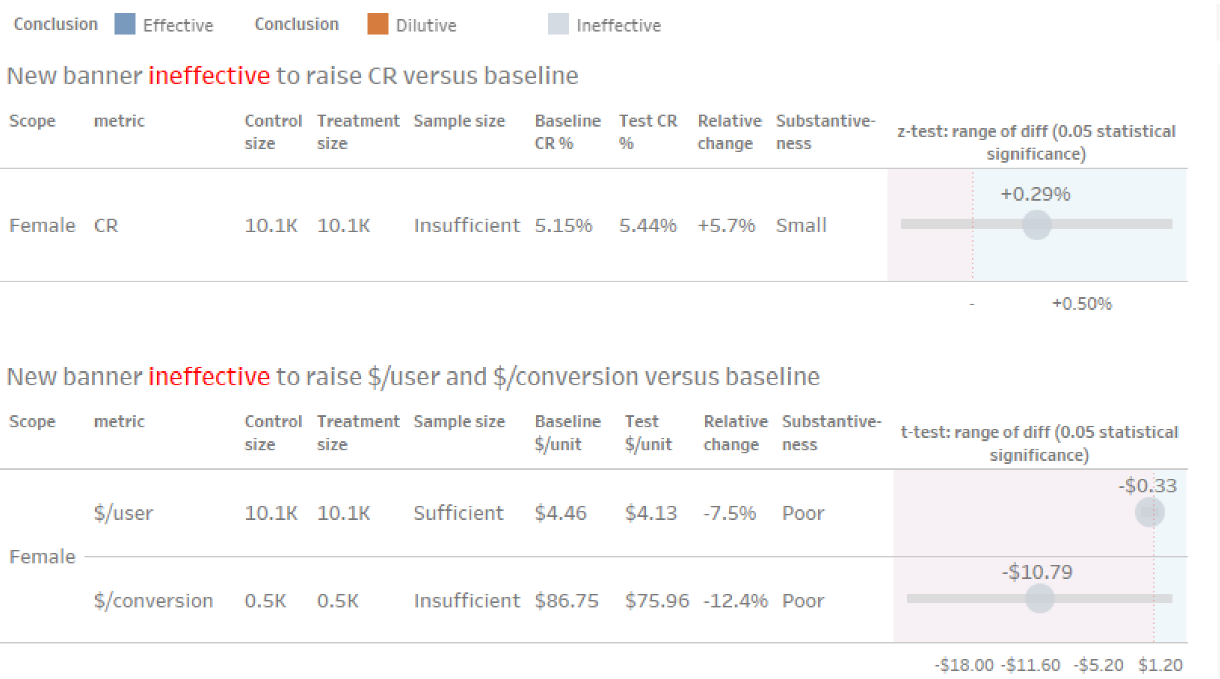
While new banner is confidently and significantly effective to raise CR, this benefit is offset by much lower monetary value being converted (as tentative evidence suggests the banner leads to diluted $/conversion).

It is **not possible to proceed to launch the experience** in its current form. Given the recommendation guidance, it is too early to **abort and not launch the experience**.

Further exploration on problems and opportunities needed to decide whether to, **perform minor iteration and then launch the experience** or **continue iterating with design of the experience and run a new A/B test**.

8. Issues and opportunities

- **Issue 1:** new banner fails to appeal to female users. Some monetary resilience as their $/conversion drops proportionally less than the aggregate.
- **Opportunity 1:** female focused program warranted. Tentative evidence below.



Conclusion 🟦 Effective   Conclusion 🟧 Dilutive   ⬜ Ineffective

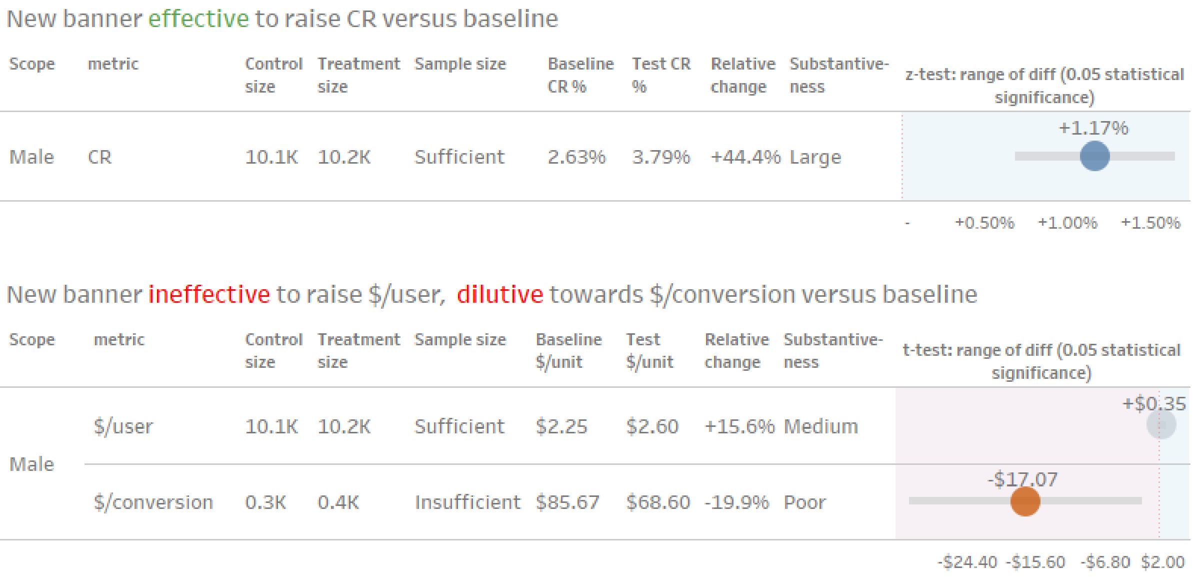
New banner ineffective to raise CR versus baseline

Scope	metric	Control size	Treatment size	Sample size	Baseline CR %	Test CR %	Relative change	Substantive-ness	z-test: range of diff (0.05 statistical significance)
Female	CR	10.1K	10.1K	Insufficient	5.15%	5.44%	+5.7%	Small	+0.29%

New banner ineffective to raise $/user and $/conversion versus baseline

Scope	metric	Control size	Treatment size	Sample size	Baseline $/unit	Test $/unit	Relative change	Substantive-ness	t-test: range of diff (0.05 statistical significance)
Female	$/user	10.1K	10.1K	Sufficient	$4.46	$4.13	-7.5%	Poor	-$0.33
	$/conversion	0.5K	0.5K	Insufficient	$86.75	$75.96	-12.4%	Poor	-$10.79

- **Issue 2:** CR boost for male users is significant but a -20% collapse in their $/conversion drives down aggregate. Baseline spending value similar to female.
- **Opportunity 2:** explore ‘premiumization’ for the banner that appeals to male users. Tentative evidence below.



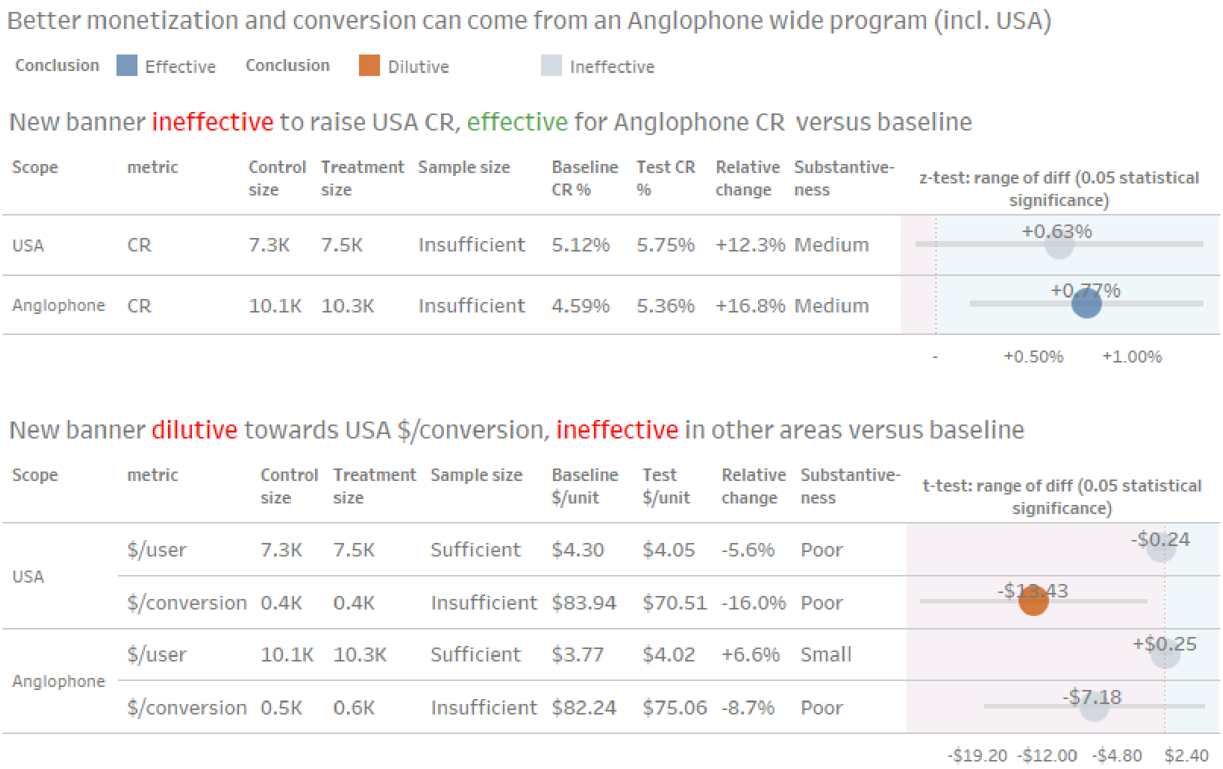
New banner effective to raise CR versus baseline

Scope	metric	Control size	Treatment size	Sample size	Baseline CR %	Test CR %	Relative change	Substantive-ness	z-test: range of diff (0.05 statistical significance)
Male	CR	10.1K	10.2K	Sufficient	2.63%	3.79%	+44.4%	Large	+1.17%

New banner ineffective to raise $/user, dilutive towards $/conversion versus baseline

Scope	metric	Control size	Treatment size	Sample size	Baseline $/unit	Test $/unit	Relative change	Substantive-ness	t-test: range of diff (0.05 statistical significance)
Male	$/user	10.1K	10.2K	Sufficient	$2.25	$2.60	+15.6%	Medium	+$0.35
	$/conversion	0.3K	0.4K	Insufficient	$85.67	$68.60	-19.9%	Poor	-$17.07

- **Issue 3:** largest market is USA. Users most engaged, but appear indifferent to new banner (albeit on cusp, p-value 0.09). Among (most) countries indifferent (CR wise), USA is the closest to 'passing grade'.
- **Opportunity 2:** Anglophone countries (USA, UK, Canada, Australia) as a bloc show statistical meaning for CR, also more resilient $/conversion. Iterating for Anglophone can be quicker and lower execution risk. USA vs Anglophone below.



Better monetization and conversion can come from an Anglophone wide program (incl. USA)

Conclusion 🟦 Effective   Conclusion 🟧 Dilutive ⬜ Ineffective

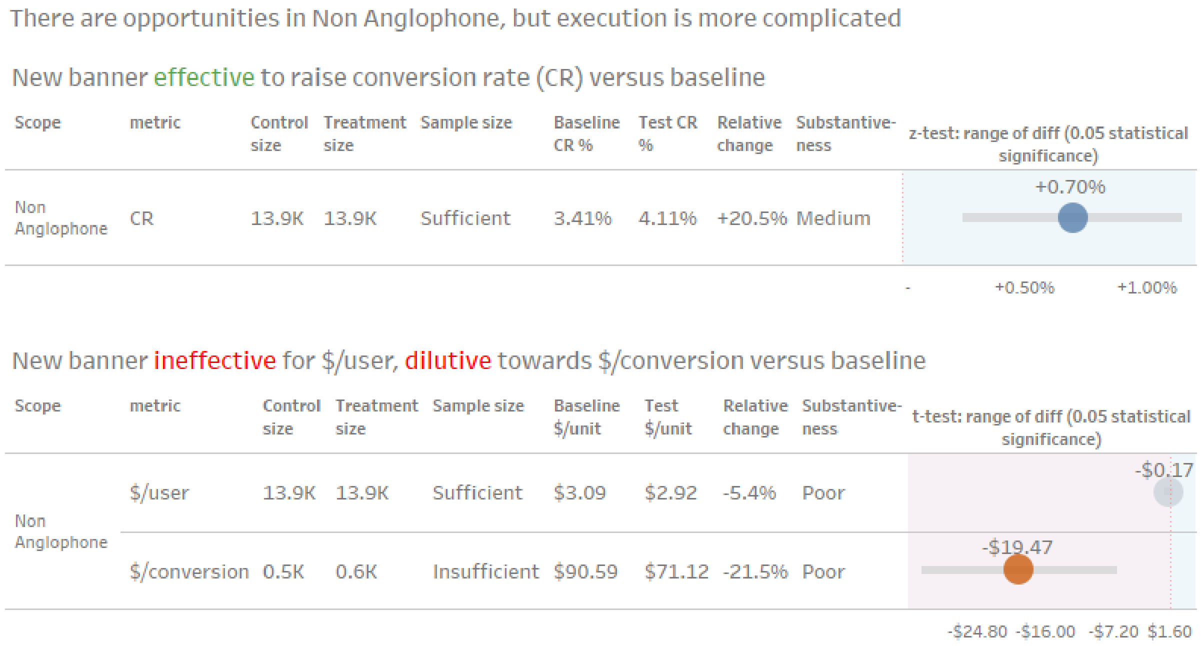
New banner ineffective to raise USA CR, effective for Anglophone CR versus baseline

Scope	metric	Control size	Treatment size	Sample size	Baseline CR %	Test CR %	Relative change	Substantive-ness	z-test: range of diff (0.05 statistical significance)
USA	CR	7.3K	7.5K	Insufficient	5.12%	5.75%	+12.3%	Medium	+0.63%
Anglophone	CR	10.1K	10.3K	Insufficient	4.59%	5.36%	+16.8%	Medium	+0.77%

New banner dilutive towards USA $/conversion, ineffective in other areas versus baseline

Scope	metric	Control size	Treatment size	Sample size	Baseline $/unit	Test $/unit	Relative change	Substantive-ness	t-test: range of diff (0.05 statistical significance)
USA	$/user	7.3K	7.5K	Sufficient	$4.30	$4.05	-5.6%	Poor	-$0.24
	$/conversion	0.4K	0.4K	Insufficient	$83.94	$70.51	-16.0%	Poor	-$13.43
Anglophone	$/user	10.1K	10.3K	Sufficient	$3.77	$4.02	+6.6%	Small	+$0.25
	$/conversion	0.5K	0.6K	Insufficient	$82.24	$75.06	-8.7%	Poor	-$7.18

- Below shows how Non Anglophone countries fare as a bloc fare, tentatively.



There are opportunities in Non Anglophone, but execution is more complicated

New banner effective to raise conversion rate (CR) versus baseline

Scope	metric	Control size	Treatment size	Sample size	Baseline CR %	Test CR %	Relative change	Substantive-ness	z-test: range of diff (0.05 statistical significance)
Non Anglophone	CR	13.9K	13.9K	Sufficient	3.41%	4.11%	+20.5%	Medium	+0.70%

New banner ineffective for $/user, dilutive towards $/conversion versus baseline

Scope	metric	Control size	Treatment size	Sample size	Baseline $/unit	Test $/unit	Relative change	Substantive-ness	t-test: range of diff (0.05 statistical significance)
Non Anglophone	$/user	13.9K	13.9K	Sufficient	$3.09	$2.92	-5.4%	Poor	-$0.17
	$/conversion	0.5K	0.6K	Insufficient	$90.59	$71.12	-21.5%	Poor	-$19.47

9. Conclusion and next steps

- **Continue iterating with design of the experience and run a new A/B test**.
- **Near term:** focus on the Anglophone bloc. Overlay fixes for female and male users as discussed.
- Design the onboarding ramp to ensure downstream sample size for $/conversion is sufficient for a good enough test (0.80 statistical power, 0.05 significance level) to detect changes observed.
- Adjust timeframe as necessary to study peak metric and novelty effects.
- Increase granularity in data captured such as transaction level product type and pre-checkout selection order to study relevant behavioral effects and unmask what appears to be ‘unaccounted noise’ in $/conversion in the test group.