Yammer - Validating A/B Test Results

The problem

When checking the results of the A/B test on July 1st, it was noticed that message posting is 50% higher in the treatment group—a huge increase in posting. The average messages sent were 2.7 for control group versus 4.1 for the test group.

We were charged to determine whether this feature is the real deal or too good to be true. The product team wants our advice about this test.

With the A/B test data that we were given, we have found some issues with the test process that need to be addressed. Plus, even with the test processing issues, the test is a failure. While it may be hard to believe, the information below will show items checked, our concerns, and our findings

Issue hypothesis

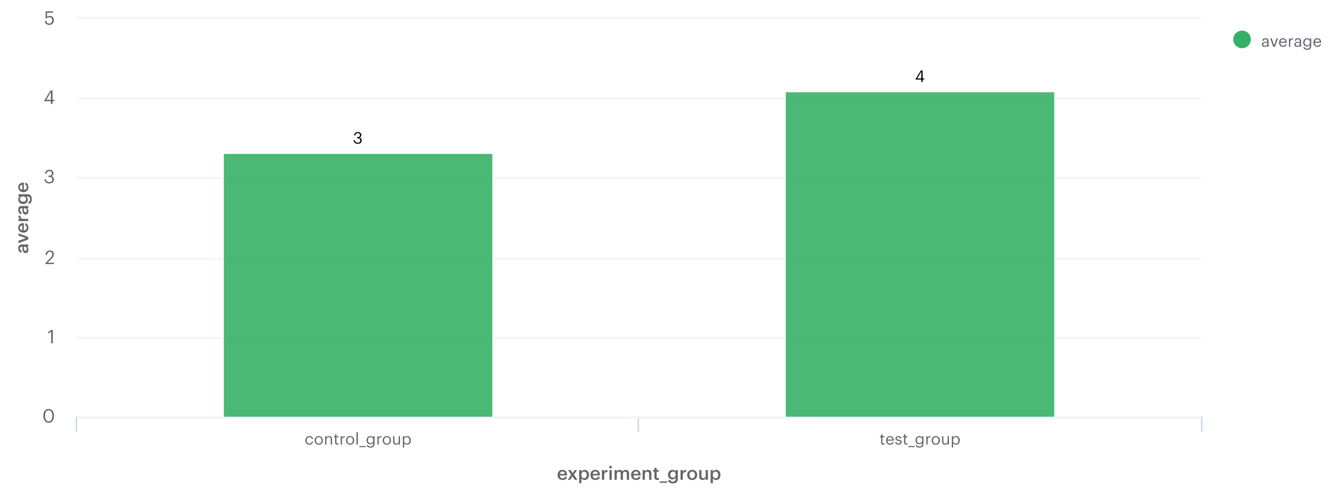
For this problem, a number of factors could explain the interesting test results. Here are a few examples:

- This metric is incorrect or irrelevant: Posting rates may not be the correct metric for measuring overall success. It describes how Yammer's customers *use* the tool, but not necessarily if they're getting value out of it. For example, while a giant "Post New Message" button would probably increase posting rates, it's likely not a great feature for Yammer. We may want to make sure the test results hold up for other metrics as well.
- The test was calculated incorrectly: A/B tests are statistical tests. People calculate results using different methods—sometimes that method is incorrect, and sometimes the arithmetic is done poorly.
- The users were treated incorrectly: Users are supposed to be assigned to test treatments randomly, but sometimes bugs interfere with this process. If users are treated incorrectly, the experiment may not actually be random.
- There is a confounding factor or interaction effect: These are the trickiest to identify. Experiment treatments could be affecting the product in some other way—for example, it could make some other feature harder to find or create incongruous mobile and desktop experiences. These changes might affect user behavior in unexpected ways, or amplify changes beyond what you would typically expect.

Validating the results

A. The number of messages sent shouldn't be the only determinant of this test's success, so we looked into a few other metrics to make sure that their outcomes were also positive. In particular, we're interested in metrics that determine if a user is getting value out of Yammer. (Yammer typically uses login frequency as a core value metric.) First, the average number of logins per user is up. This suggests that not only are users sending more messages, but they're also signing in to Yammer more.

Average Logins

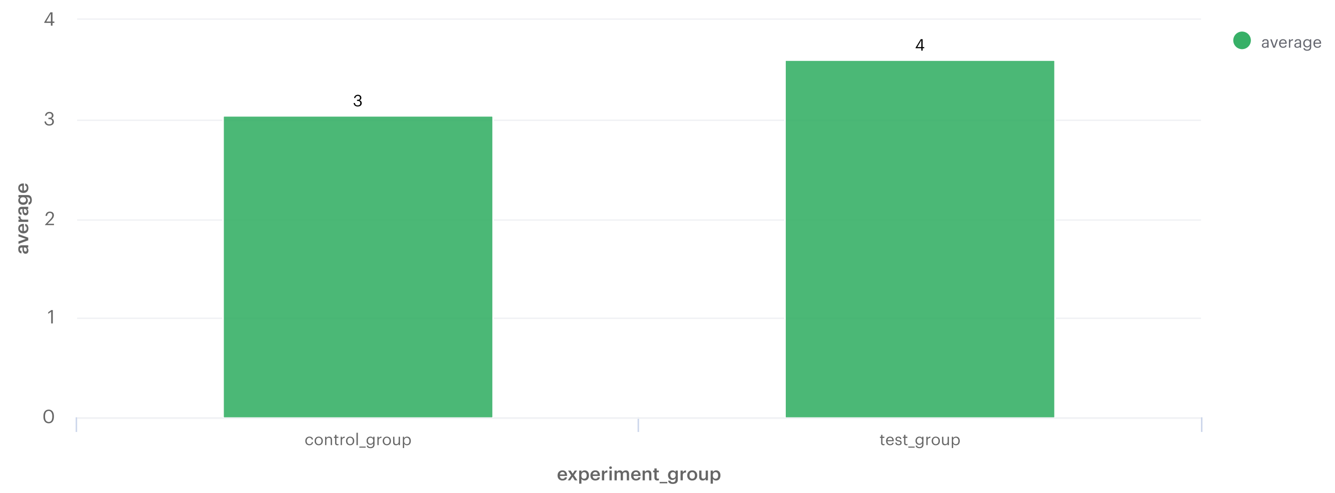


1) Avg Login Query

	experiment	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	stde
1	publisher_upda...	control_group	1746	2595	0.6728	5789	3.315...	0	0	2.5
2	publisher_upda...	test_group	849	2595	0.3272	3481	4.100...	0.7845	0.236...	3.3

Second, the users are logging in (engaging) on more days as well. If this metric was flat while logins were up, it might imply that people were logging on and logging out in quick succession, which could mean the new feature introduced a login bug. But both metric are up, so it appears that the problem is not with cherry-picking metrics.

Average Days Engaged

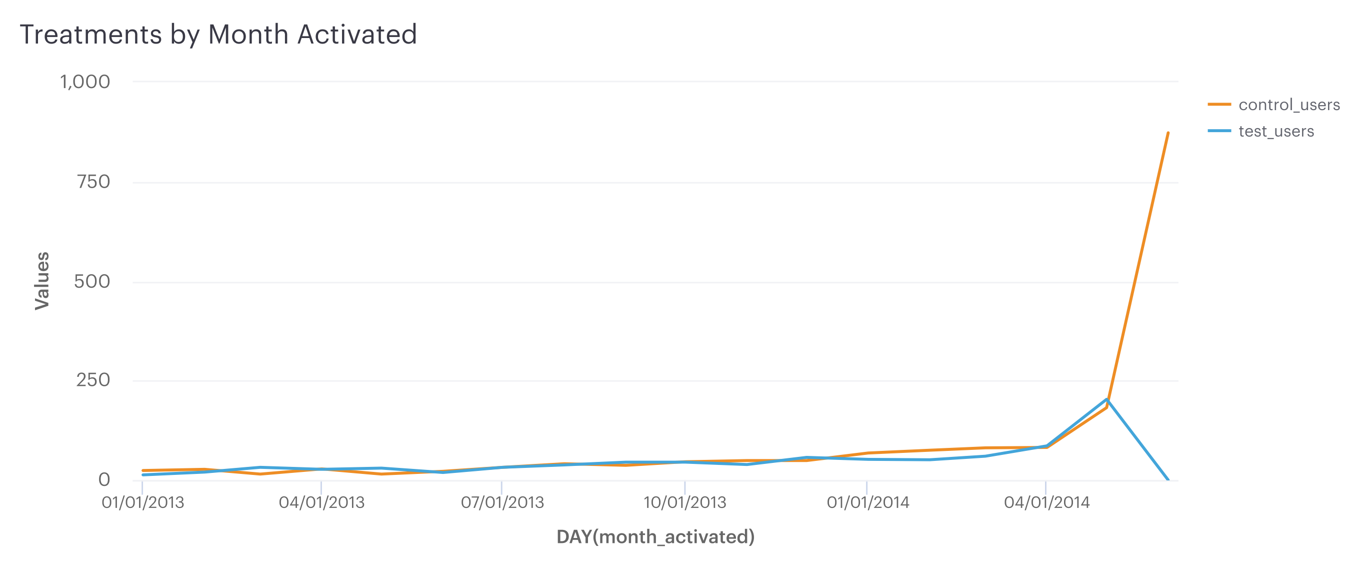


2) Avg Days Engaged Query

	experiment	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	std
1	publisher_upda...	control_group	1746	2595	0.6728	5297	3.033...	0	0	2.1
2	publisher_upda...	test_group	849	2595	0.3272	3059	3.603...	0.5693	0.187...	2.6

B. The test does suffer from a methodological error. It lumps new users and existing users into the same group, and measures the number of messages they post during the testing window. This means that a user who signed up month earlier would be considered the same way as a user who signed up a day before the test ended, even though the second user has much less time to post messages. It would make more sense to consider new and existing users separately. Not only does this make comparing magnitudes more appropriate, be it also lets you test for novelty effects. Users familiar with Yammer might try out a new feature just because it's new, temporarily boosting their overall engagement. For new users, the feature isn't "new," so they're much less likely to use it just because it's different.

Now, investigating user treatments (or splitting users out into new and existing cohorts) reveals a problem—all new users were treated in the control group.



3) Treatments by Month Activated

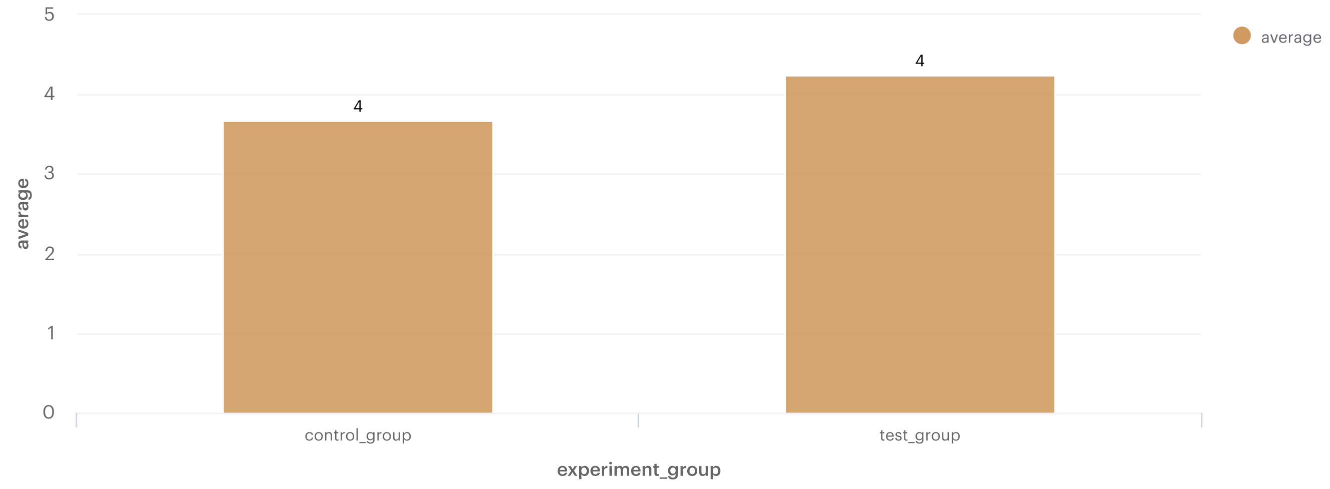
	month_activated	control_users	test_users
1	2013-01-01 00:00:00	24	13
2	2013-02-01 00:00:00	27	20
3	2013-03-01 00:00:00	15	32
4	2013-04-01 00:00:00	28	27
5	2013-05-01 00:00:00	15	30
6	2013-06-01 00:00:00	22	19
7	2013-07-01 00:00:00	32	32
8	2013-08-01 00:00:00	41	38
9	2013-09-01 00:00:00	37	45
10	2013-10-01 00:00:00	46	45
11	2013-11-01 00:00:00	49	39
12	2013-12-01 00:00:00	49	57
13	2014-01-01 00:00:00	68	52

4) Query 4

	month_activated	control_users	test_users	diff_control_minus_test
1	2014-06-01 00:00:00	873	0	873
2	2014-05-01 00:00:00	182	203	-21
3	2014-04-01 00:00:00	82	86	-4
4	2014-03-01 00:00:00	81	60	21
5	2014-02-01 00:00:00	75	51	24
6	2014-01-01 00:00:00	68	52	16
7	2013-12-01 00:00:00	49	57	-8
8	2013-11-01 00:00:00	49	39	10
9	2013-10-01 00:00:00	46	45	1
10	2013-09-01 00:00:00	37	45	-8

Looking at the Average Days Engaged (for Engagement events) before January (6 months before testing ended), the gap narrowed considerably. While this removed the issue of all the new users being treated in the control group, there are other items to look into. Plus, this treatment bug needs to be looked into before running new studies and rerunning this study.

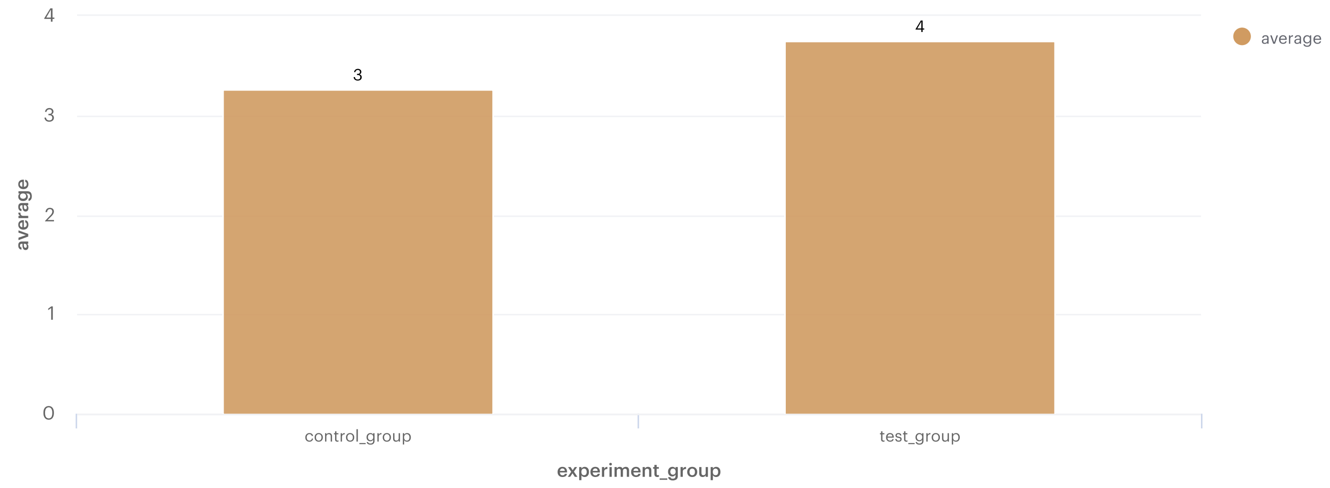
Average Logins before Jan 2014



5) Average Logins before Jan 2014

	experiment	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	std
1	publisher_upda…	control_group	385	782	0.4923	1417	3.680…	0	0	2.9
2	publisher_upda…	test_group	397	782	0.5077	1686	4.246…	0.5663	0.153…	3.3

Avg Days Engaged Query before Jan 2014



6) Avg Days Engaged Query before Jan 2014

	experiment	experiment_group	users	total_treated_users	treatment_percent	total	average	rate_difference	rate_lift	std
1	publisher_upda…	control_group	385	782	0.4923	1254	3.257…	0	0	2.3
2	publisher_upda…	test_group	397	782	0.5077	1486	3.743…	0.4859	0.149…	2.7

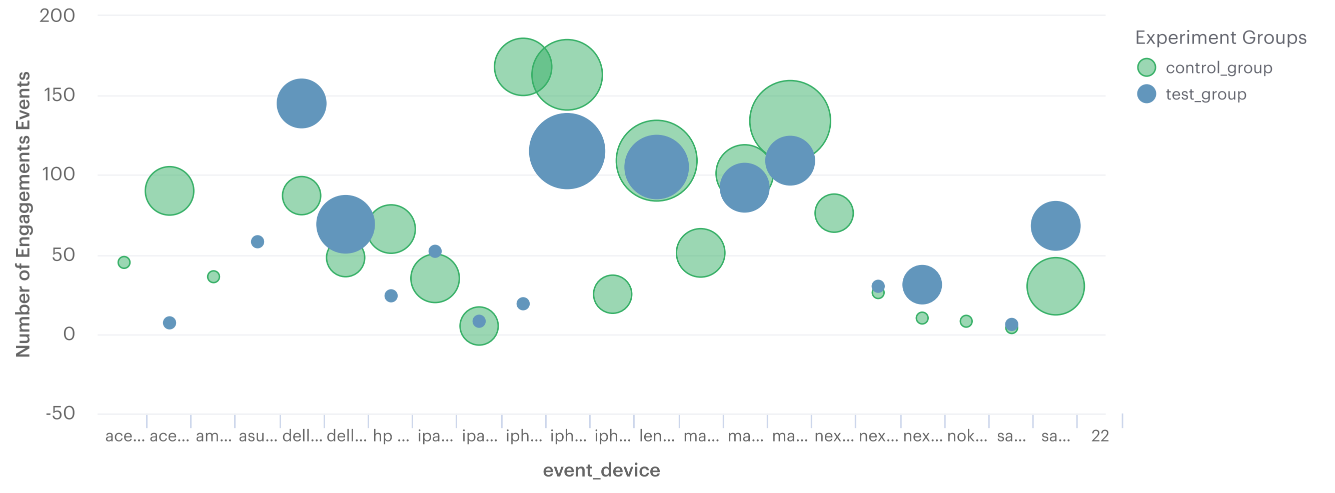
C. With the treatment issue, one should question other possible regression testing missing. One we can check now, is whether users were assigned to both the control and test groups. A quick left join shows that of none of the user_id were treated to be assigned to both groups.

7) User in Other Groups Check - Query

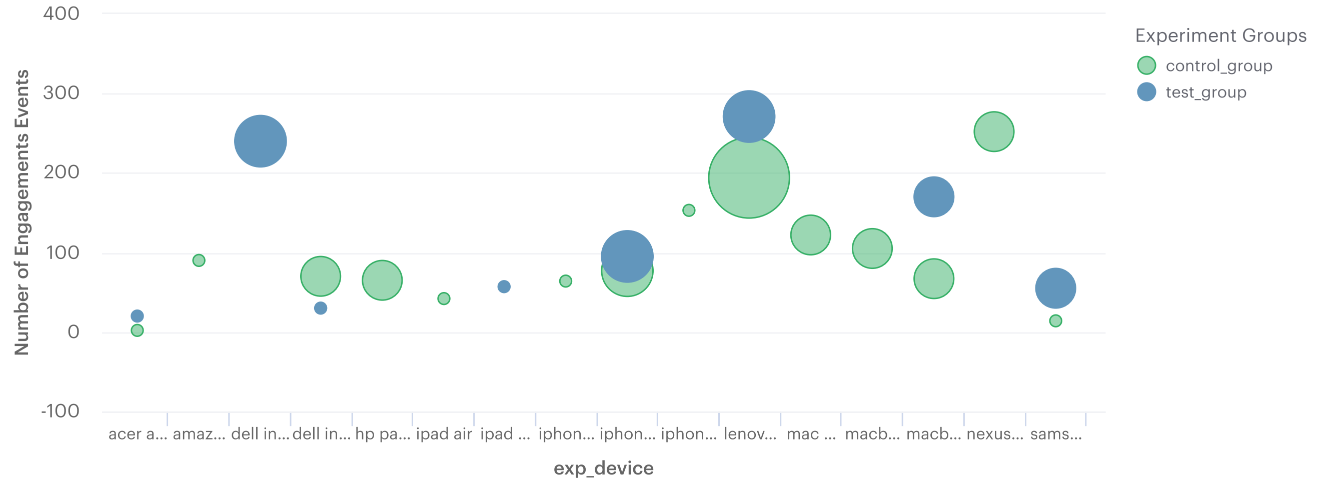
	one	two	experiment	one_grp	two_grp	one_device	two_device	one_loc	two_loc
1	4		publisher_updat…	control_grou…		lenovo thinkpad		India	
2	11		publisher_updat…	control_grou…		iphone 4s		United States	
3	19		publisher_updat…	test_group		iphone 5		Nigeria	
4	20		publisher_updat…	control_grou…		samsung galaxy s4		Italy	
5	22		publisher_updat…	test_group		lenovo thinkpad		Germany	
6	30		publisher_updat…	test_group		dell inspiron desktop		Australia	
7	49		publisher_updat…	control_grou…		hp pavilion desktop		United States	
8	59		publisher_updat…	control_grou…		macbook air		Taiwan	
9	64		publisher_updat…	control_grou…		acer aspire notebook		Turkey	
10	78		publisher_updat…	test_group		acer aspire notebook		United States	
11	83		publisher_updat…	control_grou…		amazon fire phone		United Kingdo…	
12	86		publisher_updat…	control_grou…		dell inspiron noteboo…		India	
13	98		publisher_updat…	control_grou…		dell inspiron noteboo…		Brazil	

D. Now comes the trickier part. How to find a interaction or confounding issue. The Exercise and Event tables both have device and location fields for when the user was initially treated and when the user was engaging on Yammer.

First, we looked at devices during engagement events. Where the circle size show the number of distinct users, there are a few devices with multiple users that show low numbers of engagements. Note the dots are single users. Also, at least two devices with similar number of users show a significantly lower engagement for the control group. One has to ask, is there a current issue (known or unknown) on those devices?


Event Devices for Engagement Events

Now, we need to look at the devices used during initial signup, to see if treatment could be improved to better balance the devices. While there is a lower number of devices used during initial treatment, many of the lopsidedness of user would have been help with better user balancing. Plus, both the mac mini and nexus 10 only show users in the control group during initial treatment and during engagement events.


Signup Devices for Engagement Events

The two scatter plots above and the two scatter plots below are from the same query results below.

8) Engagement Events by - Query

	one	experiment	exp_grp	exp_device	event_device	exp_loc	event_l
1	4	publisher_upda...	control_gro...	lenovo thinkpad	dell inspiron deskto...	India	India
2	4	publisher_upda...	control_gro...	lenovo thinkpad	lenovo thinkpad	India	India
3	11	publisher_upda...	control_gro...	iphone 4s	ipad air	United States	United St
4	11	publisher_upda...	control_gro...	iphone 4s	iphone 4s	United States	United St
5	19	publisher_upda...	test_group	iphone 5	iphone 5	Nigeria	Nigeria
6	19	publisher_upda...	test_group	iphone 5	lenovo thinkpad	Nigeria	Nigeria
7	19	publisher_upda...	test_group	iphone 5	nexus 7	Nigeria	Nigeria
8	20	publisher_upda...	control_gro...	samsung galaxy s4	macbook pro	Italy	Italy
9	20	publisher_upda...	control_gro...	samsung galaxy s4	samsung galaxy s4	Italy	Italy
10	22	publisher_upda...	test_group	lenovo thinkpad	asus chromebook	Germany	Germany
11	22	publisher_upda...	test_group	lenovo thinkpad	lenovo thinkpad	Germany	Germany
12	22	publisher_upda...	test_group	lenovo thinkpad	nexus 5	Germany	Germany
13	30	publisher_upda	test_group	dell inspiron deskto	dell inspiron deskto	Australia	Australia

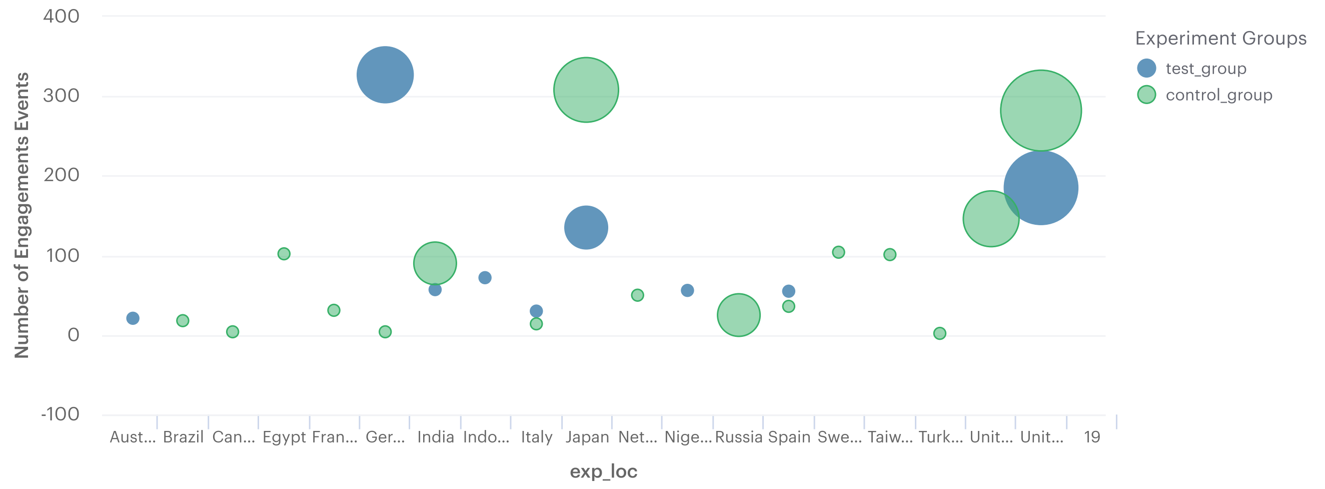
Second, we will look at the locations. With the circle sizes increasing with the number of users, a quick look at both graphs below shows that majority if not all users did not switch countries.

It looks like the new test could be hit in Germany, if the single control German with 4 engagements actually represents what Germany thinks of the current setup.

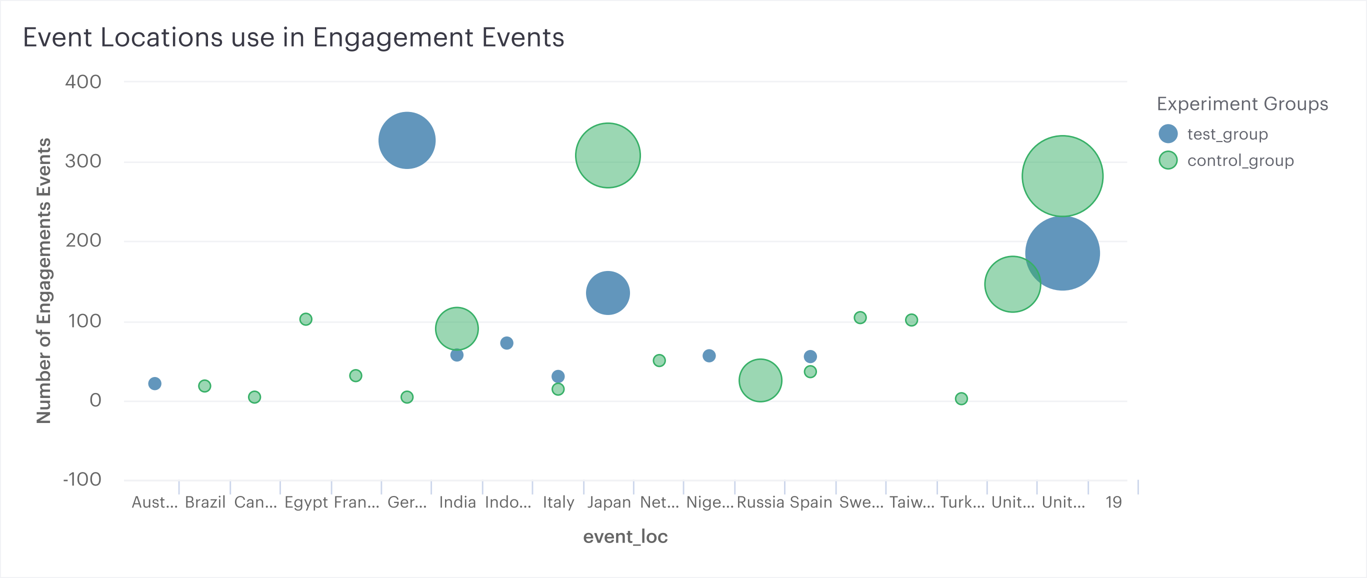
Outside of Germany every country, who were treated into both groups, preferred the current interface. Still, nothing can be made of Russia and UK low engagements with the current page, since not one was treated to the test group.

Treatment within country locations needs to be fixed so there is balance between control and test groups. In general, engagement needs to be looked into for every country outside of US and Japan.

Signup Locations for the Engagement Events



add



Conclusion

The treatment to assign users to either the control group or test group need to fixed. Balance within locations and between devises need to be introduced. This may be part of the last day issue, where 872 users were assigned to control group.

Even with the above in mind, the test change looks to be a failure. There is a possibility that German may like it, but their one control user should be considered an outlier. Without any control users, we do not know if the German test results are positive or negative. Most likely Germans may just like and use Yammer more than others. Plus, additional research may find that some of the 872 new users are part of Japan, US, and India control groups. This will mean that the control group may have less user for the same number of engagements. That would make the test interface results look a little worse than the Event Location graph currently seems.