CONSUMER ACQUISITION

2021 UPDATE

Facebook Creative Testing

Why Is The Control Video So Hard To Beat?

Table of Contents

Section 1

Creative Testing

Our Unique Way

Let us set the stage for how and why we have been doing creative testing in a unique way. We test a lot of creative. In fact, we produce and test more than 100,000 videos and images yearly for our clients, and we have performed over 10,000 A/B and multivariate tests on Facebook, Google, TikTok and Snap.

We focus on these verticals: gaming, e-commerce, entertainment, automotive, D2C, financial services, and lead generation. When we test, our goal is to compare new concepts versus the winning video (control) to see if the challenger can outperform the champion. Why? If you cannot outperform the best ad in a portfolio, you will lose money running the second or third place ads.

While we have not tested our process beyond the verticals, we have managed over $3 billion in paid social ad spend and want to share what we've learned. Our testing process is designed to save both time and money by killing losing creatives quickly and to significantly reduce non-converting spend. Our process will generate both false negatives and false positives. We typically allow our tests to run between 2-7 days to provide enough time to gather data without requiring the capital and time required to reach statistical significance (StatSig). We always run our tests using our software AdRules via the Facebook API. Our insights are specific to the above scenarios, not a representation of how all testing on Facebook's platform operates. In cases, it is valuable to retain learning without obstructing ad delivery.

To be clear, our process is not the Facebook best practice of running a split test and allowing the algorithm to reach statistical significance (StatSig) which then moves the ad set out of the learning phase and into the optimized phase. The insights we have drawn are specific to the scenarios we outline here. They are not a representation of how all testing on Facebook's platform operates. In some cases, it is valuable to have old creative retain learning to seamlessly A/B test without obstruct- ing ad delivery.

Statistical Significance vs Cost-Effective

Let us take a closer look at the cost aspect of creative testing.

In classic testing, you need a 95% confidence rate to declare a winner, exit the learning phase, and reach StatSig. That's nice to have, but getting a 95% confidence rate for in-app purchases may end up costing you $20,000 per creative variation.

Why so expensive?

Here is an example scenario: To reach a 95% confidence level, you'll need about 100 purchases. With a 1% purchase rate (which is typical for gaming apps), and a $200 cost per purchase, you will end up spending $20,000 for each variation in order to accrue enough data for that 95% confidence rate. There are not a lot of advertisers who can afford to spend $20,000 per variation, especially if 95% of new creative fails to beat the control. In this example, with a cost of $20,000 per variation and a 95% failure rate, it would cost $400,000 just to find a new control. That is because we would have to test 20 variations to find the winner, and – as mentioned earlier – testing each variation costs about $20,000.

Stat Sig Is Expensive	Non-Stat Sig: Cheap IPM Testing
95% Stat Sig: Game Averages	95% Stat Sig: Game Averages
1% Purchase Rate	0.5% Install Rate
100 Purchases	100 installs
$200 $/Purchase	$2.00 $/Install
$20,000 / Creative	$200 / Creative ($20 Intl') **BOOM**

So, what to do?

To avoid such high testing costs, we move the conversion event we are targeting up, or towards the beginning of the sales funnel. For mobile apps, instead of optimizing for purchases, we optimize for impressions per install (IPM). For web- sites, we would optimize for an impression to top-funnel conversion rate. Again, this is not a Facebook recommended best practice. This is our own methodology, designed to allow advertisers to find new, top-performing creative in the most cost-efficient and reliable way.

IPM Testing Is Cost-Effective

A concern with our process is that ads with high CTRs and high conversion rates for top-funnel events may not be true winners for down-funnel conversions and ROI / ROAS. But while there is a risk of identifying false positives and negatives with this method, we would rather take that risk than spend the time and expense of optimizing for StatSig bottom-funnel metrics.

To us, it is more efficient to optimize for IPMs vs. purchases. Most importantly, it means you can run tests for less money per variation because you are optimizing towards installs vs purchases. For many advertisers, that alone can make more testing financially viable. $200 testing cost per variation versus $20,000 testing cost per variation can mean the difference between being able to do a couple of tests versus having an ongoing, robust testing program.

We do not just test a lot of new creative ideas. We also test our creative testing methodology. That might sound a little "meta," but it is essential for us to validate and challenge our assumptions and results. When we choose a winning ad out of a pack of competing ads, we would like to know that we have made a good decision.

Because the outcomes of our tests have consequences – sometimes big consequences – we test our testing process. We question our testing methodology and the assumptions that shape it. This is because the outcome of every test shapes our creative strategy going forward. So, making the wrong call on a test of new creative concepts does not just mean we kill the concepts that did not work. Rather, it means our whole creative strategy pivots away from those concepts. If we are wrong, that could have significant consequences.

Section 2

How We Test Creative Now

How We Have Been Testing Creative Until Now

When testing creative we typically would test three to six videos along with a control video using Facebook's split test feature. We would show these ads to broad or 5-10% LALs (Lookalike) audiences, and restrict distribution to the Facebook newsfeed only, Android only and we would use mobile app install bidding (MAI) to get about 100-250 installs.

If one of those new "challenger" ads beat the control video's IPM or came within 10%-15% of its performance, we would launch those potential new winning videos into the ad sets with the control video and let them fight it out to generate ROAS.

We have seen hints of what we're about to describe across numerous ad accounts and have confirmed with 7-figure spending advertisers that they have seen the same thing. But for purposes of explanation, let us focus on one client of ours and how their ads performed in creative tests.

In November and December 2019, we produced +60 new video concepts for this client. All of them failed to beat the control video's IPM. This struck us as odd, and it was statistically impossible. We expected to generate a new winner 5% of the time or 1 out of 20 videos – so 3 winners. Since we felt confident in our creative ideas, we decided to look deeper into our testing methods.

The traditional testing methodology includes the idea of testing a testing system or an A/A test. A/A tests are like A/B tests, but instead of testing multiple creatives, you test the same creative in each "slot" of the test.

If your testing system/platform is working as expected, all "variations", should produce similar results assuming you get close to statistical significance. If your A/A test results are very different, and the testing platform/methodology concludes that one variation or another significantly outperforms or underperforms compared to the other variations, there could be an issue with the testing method or quantity of data gathered.

First A/A test of video creative

Here is how we set up an A/A test to validate our non-standard approach to Facebook testing. The purpose of this test was to understand if Facebook maintains a creative history for the control and thus gives the control a performance boost making it very difficult to beat – if you do not allow it to exit the learning phase and reach statistical relevance.

We copied the control video four times and added one black pixel in different locations in each of the new "variations." This allowed us to run what would look like the same video to humans but would be different videos in the eyes of the testing platform. The goal was to get Facebook to assign new hash IDs for each cloned video and then test them all together and observe their IPMs.

These are the ads we ran… except we did not run the daschund; I have replaced the actual ads with cute dogs to avoid disclosing the advertiser's identity. IPMs for each ad in the far right of the image.
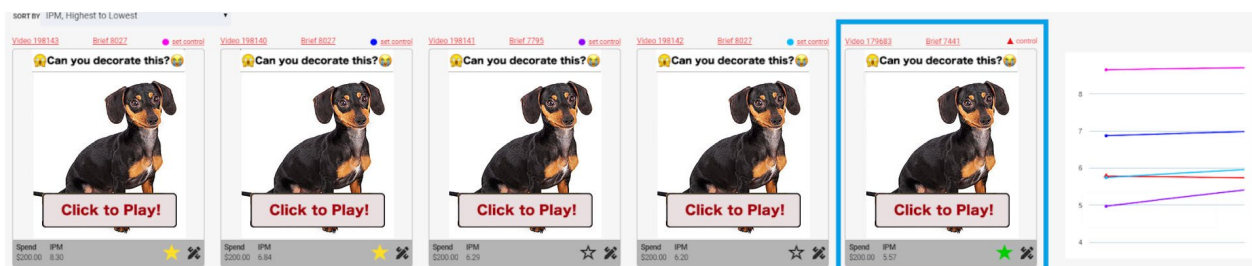
Things to note here:

The far-right ad (in the blue square) is the control.

All the other ads are clones of the control with one black pixel added.

The far-left ad/clone outperformed the control by 149%. As described earlier, a difference like that should not happen. If the platform was truly variation agnostic, BUT – to save money, we did not follow best practices to allow the ad set(s) to exit the learning phase.

We ran this test for only 100 installs. Which is, our standard operating procedure for creative testing.

Once we completed our first test to 100 installs, we paused the campaign to analyze the results. Then we turned the campaign back on to scale up to 500 installs to get closer to statistical significance. We wanted to see if more data would result in IPM normalization (in other words if the test results would settle back down to more even performance across the variations). However, the results of the second test remained the same. Note: the ad set(s) did not exit the learning phase and we did not follow Facebook's best practice.

The results of this first test, while not statistically significant, were surprisingly enough to merit additional tests. So, we tested on!

Second A/A test of video creative

For our second test, we ran the six videos shown below. Four of them were controls with different headers; two of them were new concepts that were very similar to the control. Again, we didn't run the daschund; they've been inserted to protect the advertiser's identity and to offer you cuteness!

The IPMs for all ads ranged between 7-11 – even the new ads that did not share a thumbnail with the control. IPMs for each ad in the far right of the image.
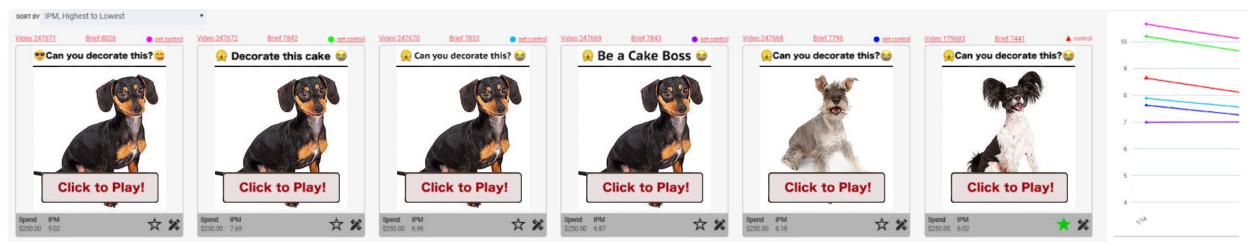
Third A/A test of video creative

Next, we tested six videos: one control and five visually similar variations to the control but one very different from a human. IPMs ranged between 5-10. IPMs for each ad in the far right of the image.
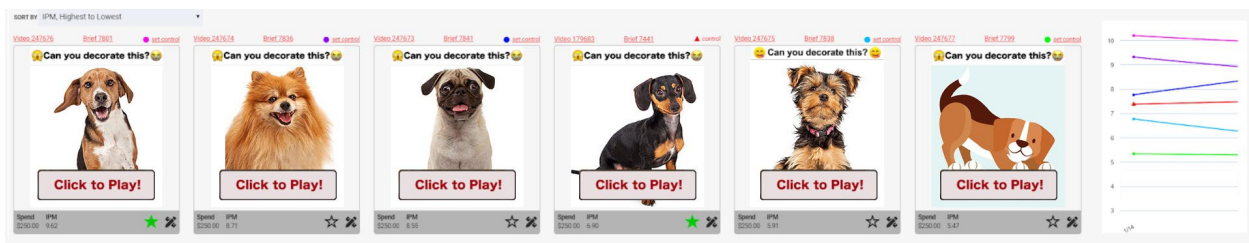
Fourth A/A test of video creative

This was when we had our "ah-ha!" moment. We tested six very different video concepts: the one control video and five brand new ideas, all of which were visually very different from the control video and did not share the same thumbnail.

The control's IPM was consistent in the 8-9 range, but the IPMs for the new visual concepts ranged between 0-2. IPMs for each ad in the far right of the image.
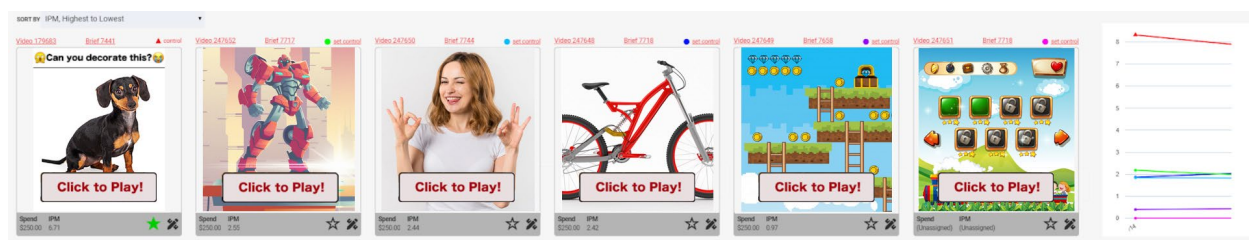
Here are our impressions from the above tests

Facebook's split-tests maintains creative history for the control video. This gives the control advantage with our non-statistically relevant, non-standard best practice of IPM testing.

We are unclear if Facebook can group variations with a similar look and feel to the control. If it can, similar-looking ads could also start with a higher IPM based on influence from the control — or perhaps similar thumbnails influence non-statistically relevant IPM.

Creative concepts that are visually very different from the control appear to not share a creative history. IPMs for these variations are independent of the control.

It appears that new, "out of the box" visual concepts vs the control may require more impressions to quantify their performance.

Our IPM testing methodology appears to be valid if we do NOT use a control video as the benchmark for winning.

IMP Testing Summary

Here are the line graphs from the second, third, and fourth tests.
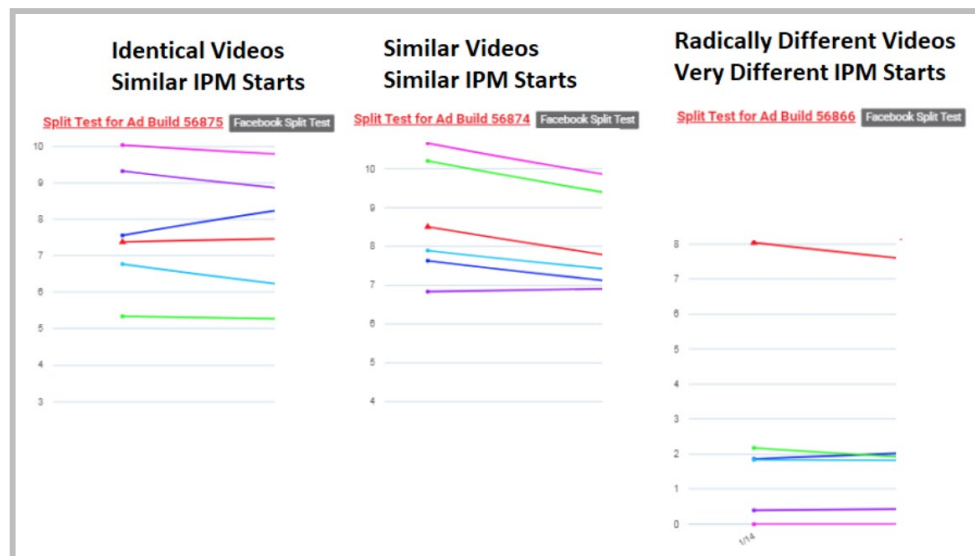
In the next section, we will explain what we think all this means, and how to work with the implications of these findings in your campaigns.

Creative Testing Recommendations 2.0

And here is what we think they mean:

Creative Testing Recommendations 2.0:

Because of the A/A test findings that we outlined in the past section, we decided to re-test our IPM winners while excluding the control video so we could determine if we have been killing potential winners.

Drawing from what we learned when we re-ran those tests, we updated our testing methodology. This is not the first time we have done this – our testing methodology is constantly evolving – but some significant changes had to be made once our understanding of Facebook's testing algorithm evolved.

For Q1 2021, we recommend one of the three-phase testing plans outlined below, depending on whether your app monetizes with purchases or advertising.

But before we dive into that, there is one major issue we know you are curious about, and it applies to both IAP and IAA advertising: iOS 14.

Facebook's Creative Recommendations for Game Marketers in 2021

Facebook Gaming recently published a new report "Games Marketing Insights for 2021" in which they detail the impact of the COVID-19 pandemic over the past year on mobile gaming. By comparing cohorts of existing gamers to the millions of new mobile gamers that emerged in 2020, Facebook uncovered some interesting insights for mobile marketers.

Most notably, Facebook discovered that "Gamers want to see gameplay in ads."

When it comes to gaming ads, both cohorts of existing and new gamers have similar preferences: everyone prefers to see the main gameplay and story. Specifically, gamers want to see ads that showcase the game's art style, the characters, as well as high scores that players can achieve.

Mobile game marketers would do well to adopt a creative approach focused on storytelling, taking into consideration that people like to see gameplay in ads. By telling rich stories through immersive experiences and new formats, you can create familiarity, which we know drives discovery. Consider the creative elements in your ads with other creative means of engaging players such as streamer partnerships which have the potential to influence prospective gamers.

What happens when those iOS 14 privacy filters go up?

It is pretty simple: We'll switch over to Android to do creative testing.

This is not even all that huge a change. We have generally tried to test on Android even before because it is less expensive. The only time we do not test on Android is when the client doesn't have an Android app (which is very rare) or they're just dead set against it.

The best news is that Android wins translate to iOS. So, in some ways, the iOS update is just cementing what we've been doing before. Also, CBO and business as usual winners translate over to AAA. So the iOS 14 change will have some effects, but it may not be as catastrophic as it first appeared.

Our 3-Step Creative Testing Process for IAP (In-App-Purchases)

1

IPM TEST
- New campaign for testing
- NO control video
- 1 ad per ad set
- FB split test, 3-6 creatives
- Drive 100 installs
- Either operating systems
- Newsfeed or Fan
- Winner(s) move to #2 for IAP titles

2

INITIAL ROAS
- New campaign AEO/VO
- NO control video
- 1 ad set, multiple ads
- Budget $500/day 7 days
- Either operating systems
- All placements
- 2-3 top audiences
- Winner(s) move to #3

3

ROAS SCALE
- Strong CBO Campaign (AEO / VO)
- NO control video
- New ad set
- Exit the learning phase
- Winner(s) back to Client for UA teams to scale

Phase 1: IPM Test

- No control videos

- Create a new split test campaign using 3~6 new creatives (no control).

    - Setup campaign structure for basic App Install (No event optimization or value optimization)

    - Spend an equal amount on each creative. Ex: One ad per ad set.

    - Budget for at least 100 installs per creative
        - $200~$400 spend per ad is recommended (based on a CPI of $2-$4) if T1 English-speaking country
        - $20~$40 spend per ad/adset testing in India (based on $0.20-$0.40 CPI)

    - US Phase 1 testing.
        - 10-15% LAL with a seed audience similar to past 90-day installers, or past 90-day payers.

    - Non-US Phase 1 testing.
        - Use broad targeting & English speakers only
        - If not available in India, try other English-speaking countries with lower CPMs than U.S. and similar results. Ex: ZA, CA, IE, AU, PH, etc.

    - Use the OS (iOS or Android) you intend to scale in production

    - Use one body text

    - Headline is optional

    - FB Newsfeed or Facebook Audience Networking placement only (not both and not auto placements)

    - Be sure the winner has 100+ installs (50 installs acceptable in high CPI scenarios)
        - 100 installs: 70% confidence with 5% margin of error
        - 160 installs: 80% confidence with 5% margin of error
        - 270 installs: 90% confidence with 5% margin of error
        - IAP Titles: kill losers, top 1~3 winners go to phase 2
        - IAA Titles: kill losers, allow top 1~3 "possible winners" to exit the learning phase and then put into "the Control's" campaign

Which Creatives Move from Phase 1 > Phase 2?
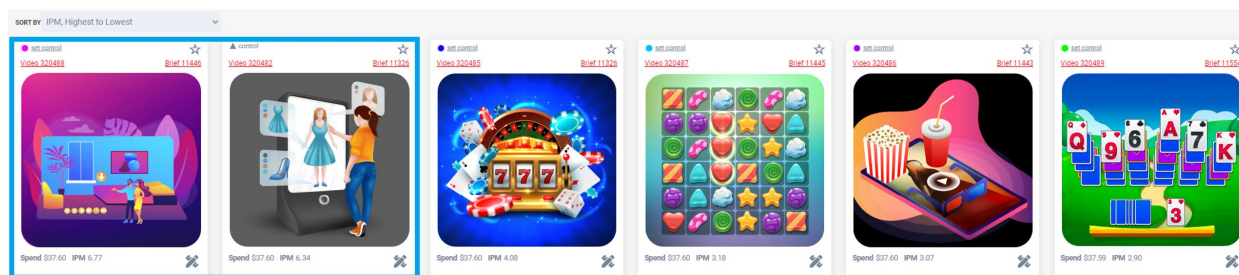
How To Pick A Phase 1 IPM Winner?

    - IPMs may range broadly or be clumped together

    - Goal: kill obvious losers and test remaining ads in phase 2

    - Ads (blue) have IPMs 6.77 & 6.34, move to phase 2

• If all ads are very close (e.g., within 5%), increase the budget
• IAA (in-app ads titles) you may need more LTV data before scaling

Phase 2: Initial ROAS

• No control videos

• Create a new campaign with AEO or VO optimization

• Place all creatives into a single ad set (Multi Ads Per Ad Set)

• Use IPM winner(s) from Phase 1 (you can combine winners from multiple Phase 1 tests into a single Phase 2 test)

• OS – Android or iOS. 5-10% LALs from top seeds (purchases, frequent users + purchase) + Auto Placements

• Testing can be done at a lower cost if you wish to run this campaign in other countries where ROAS is similar or higher but CPMs are much lower compared to the US – ie., South Africa, Ireland, Canada, etc.

• Lifetime budget $3,500-$4,900 or daily budgets of $500-$750 over 4-6 day (depending on your $/purchase).

- **WARNING!** Skipping this step is highly likely to result in one of the following scenarios:
    - Challenger immediately kills the champion/control but hasn't achieved enough statistical relevance or exited the learning phase and therefore the sustained ROAS/KPI may not be sustained.

    - Champion/control video has a lot more statistical history and relevance and most likely has exited the learning phases and may immediately kill the challenger before it has a chance to get enough data to properly fight for ROAS.

- **The latest change to our Phase 2 strategy:** we have become more focused on retention than we were a few months ago. Before, we were just looking at pure ROAS. Now, we pay attention to retention information as it comes in on Day 1 to Day 3 or so. This is usually possible, because we'll run Phase 2 ads for about 5 days, and that's worth weighing in for which pieces of creative you take to Phase 3.

Phase 3: ROAS Scale

- No control videos

- Use strong CBO campaign

- Choose winner(s) from Phase 2 with good/decent ROAS
    - You have proven the ad has great IPM and "can monetize"
    - To win this phase, it must hit KPIs (D7 ROAS, etc.)

- Create a copy of an existing ad set
    - Delete old ads and replace them with your Phase 2 winner(s)
    - Allows new ads to spend in a competitive environment

- Then, create a new ad set, roll it out towards target audiences with solid ROAS / KPIs

- CBO controls budgets between ad sets with control creatives and ad sets with new creative winners.
    - Intervene with ad set min/max spend control only if new creatives do not receive spend from CBO.

- Require challenger to exit the learning phase before moving to challenge the control "Gladiator" video.

- Once the challenger has exited the learning phase, allow CBO to change budget distribution between challenger and champion.

Our 3-Step Creative Testing Process for IAA (In-App-Advertising)

It is pretty simple: We'll switch over to Android to do creative testing.

This is not even all that huge a change. We have generally tried to test on Android even before because it is less expensive. The only time we do not test on Android is when the client doesn't have an Android app (which is very rare) or they're just dead set against it.

The best news is that Android wins translate to iOS. So, in some ways, the iOS update is just cementing what we've been doing before. Also, CBO and business as usual winners translate over to AAA. So the iOS 14 change will have some effects, but it may not be as catastrophic as it first appeared.

1 — IPM TEST
- New campaign for testing
- NO control video
- 1 ad per ad set
- FB split test, 3-6 creatives
- Drive 100 installs
- Either operating systems
- Newsfeed or Fan
- Winner(s) move to #2 for IAP titles

2 — INITIAL RPM
- New campaign AEO (not VO)
- Chose geos based on RPM (high revenue) and start there
- 1 ad set, multiple ads
- Target achieve level
- Budget $250
- Auto Placements
- 2-3 Top Audiences

3 — ROAS SCALE
- Not ROAS for IAA
- Pay attention to Retention

Phase 1: IPM Test

- We use the same approach for IAA ads as we do for IAP ads in Phase 1.

Phase 2: Initial RPM

- Look at your MMP. Find whichever countries/regions/geos are paying out a good RPM, and just go with those.

- No control videos. Multiple ads per ad set. We call this "the gladiator battle."

- Create a new campaign with AEO (not VO), but instead of the event being a purchase, have it be an event that could only occur after someone has played the game for a long time like they've achieved a certain level.

- For AEO, with a nonpurchase event, you can lower your budget down to even $250 per day.

- Auto placements are okay if that's what you typically use for the account. Generally, just use the placements you already know will work. No need to reinvent the wheel here. Just run your default best setup.

- For audiences, just run your top two to three audiences. If you are concerned about budget, stack your audiences so you only have one ad set. This will get the ad set out of the learning phase faster and save you some money. (Keep in mind that you may not get out of the Learning Phase at all sometimes in Phase 2 testing.)

Phase 3: ROAS Scale

- For IAA, there really is not a Phase 3, but our recommendations for what to do at this stage depend on your budget:

    - If your budget is small, you are not going to know performance for several weeks and so you might as well just roll your best-performing ads out into production.

    - If your budget is large, do a scaled-down Phase 3 structure as we suggest for IAP advertising. This is especially important if you have control ads that are still doing well. Roll strong Phase 2 performers whenever you need a win.

Automation on Facebook AAA and Google UAC

We are running a lot of ads on Facebook and Google's versions of automated ads. We do not tend to test in either of these campaign types (more on that below), but they're an essential part of our work now, and they need to be run very differently than standard campaigns.

As with standard campaigns, IDFA is expected to have some major impacts, like:

- Value Optimization as we've been using it is over
- It's possible Facebook iOS 14 ad accounts will have a campaign and ad limits
- Google hasn't told us what they expect for UAC
- We should focus on upper-funnel events in Q1 so we can proxy future behavior
- Facebook's audience network is going to be substantially smaller
- We will shift into testing on Android before we run on iOS

So that is what we expect once IDFA goes live. For now, here is how we are managing Facebook's AAA and Google's UAC.

Facebook's AAA

ADVANTAGES

- You get creative performance reporting on images, videos, text, and you can get that tied to standard events. This reporting persists (Google's doesn't). We do all of our testing on Facebook because of the granular controls we have, both for bidding, creative reporting, and targeting.

- You have a choice between an auto bid and a bid cap.

- There are fewer restrictions on the text. You can use CAPS, exclamation points, emojis, etc.

- Flexibility with creative assets: In Facebook, you can basically do whatever you want with 50 assets, and the reporting will persist. So, if you want to run 49 videos and one image, great. (Note: We also suspect that Facebook will soon give advertisers even more flexibility with how many creative assets they can use. They know that creative testing is the best competitive advantage left for advertisers, and they want to give us the tools to aggressively optimize our ads.)

- You can assume there is going to be a CPM discount and a CPI discount. It's just like CBO – whenever you use Facebook solutions that they are encouraging, the platform will give you a discount.

- Android and iOS performance can be comparable to standard campaigns.

- 4 x 5 and square aspect ratios tend to do best on Facebook, though in some placements, full portrait (9 x 16) is better. Be careful about Facebook allowing you to run creative assets that don't entirely fit on given placements. Facebook will just shave off the top and bottom of your ads in some cases, which could mean they are chopping off the ad's CTA or other critical information.

Google UAC

ADVANTAGES

- Flexibility in campaign bidding structure.

- You can have multiple campaigns running in any geo. You can have different CPAs, campaigns just running for installs, whatever.

- Better transparency for performance. You can have a 1:1 event reconciling with MMP data.

- You can see the performance by traffic source.

- You can have multiple Ad Groups, and so you can put different creative approaches into separate "buckets" and see how their performance compares. This also means you can run seasonal or ad hoc events without disturbing evergreen campaigns.

- Being able to create multiple Ad Groups also means you can test a lot more creative than you can on Facebook. Google does cap your assets at 20 videos and 10 images, but if you need more you can just create more Ad Groups.

- Portrait (9 x 16) and landscape (16 x 9) aspect ratios are the best bet right now for videos on UAC.

CHALLENGES

- In Google, you can have up to 20 videos and 10 images. Those are hard caps; you cannot swap out and have even 21 videos and 9 images. But if you swap those assets out to add new assets, you will not be able to see the stats for those creative assets. It is possible to get the stats back if you re-add the videos. So, the data isn't lost. It's just not shown.

- Reporting on creative performance is harder to get.

- You cannot control bid settings (like an auto bid and bid cap).

- You cannot control where your traffic is coming from, or how it is allocated throughout your campaigns, Ad Groups, and ads. This means you cannot really do a proper split-test. You can have one creative in one Ad Group, and then create as many Ad Groups as creative variations, but that is not a proper split-test.

- Google is more restrictive about the text. It will not allow CAPS, exclamation points, emojis, etc.

- There are no standard campaign options on Google. There's only UAC.

- iOS does poorly on Google UAC. Android does fine. Compare this to Facebook, where both platforms tend to perform about the same.

Creative Sizes Best Practice

To maximize creative testing, run the best aspect ratio and media types across Facebook, Google App Campaigns, TikTok and Snap. However, as the failure rate of new creative is 95% you will also want to minimize the amount of sizes you produce until you have uncovered a fresh winning creative. Below is our recommendation to maximize distribution and while minimizing creative production.

Facebook

Video Required		16:9	4:5
9:16	**1:1**	**16:9**	**4:5**
Vertical	Square	Lansdcape	Portrait
1080 x 1920	1080 x 1080	1920 x 1080	1080 x 1350

Instagram

Video Required		16:9	4:5
9:16	**1:1**	**16:9**	**4:5**
Vertical	Square	Lansdcape	Portrait
1080 x 1920	1080 x 1080	1920 x 1080	1080 x 1350

Google

Video Required		1.91:1	Image	2:3	3:2	3:4	4:3
9:16	**1:1**	**1.91:1**	**1.91:1**	**2:3**	**3:2**	**3:4**	**4:3**
Vertical	Square	Landscape	1200 x 628	320 x 480	480 x 320	768 x 1024	1024 x 768
1080 x 1920	1080 x 1080	1200 x 628					

TikTok

Video Required	Image
9:16	**1:1**
Vertical	Square
1080 x 1920	1080 x 1080

Snap

Video Required
9:16
Vertical
1080 x 1920

So, what does all this mean for User Acquisition Managers and Agencies in the Near Future?

As we have written about before, it is promising for anyone who focuses on creative optimization. That is absolutely the thing we will have the most control over going forward. So, pivot away from the hands-on campaign management that the algorithms can now actually do better than humans can. Focus on competitive analysis, creative optimization, and creative strategy.

CONSUMER ACQUISITION