A framework for improving advertising creative using digital measurement

Lara Andrews, Adetunji Olojede, Robert Moakler, Nikhil Nawathe, Mark Zhou

Abstract

While creative is thought to be an important factor in advertising success, there is no clear industry consensus on how to quantify this impact or use it to identify creative best practice. This paper reviews industry literature that has broached this topic to date and provides a three-stage framework for how to approach the problem of creative testing and measurement. Firstly, the authors advocate for the identification of a creative attribution metric, to isolate the creative impact from other factors that impact campaign. Secondly, it serves to deconstruct creative into its component parts, to reduce complexity by limiting the levels of variation the researcher needs to account for. Domain knowledge should be applied to make decisions about which creative features might be both impactful and measurable. Finally, it is recommended that researchers uncover causal relationships between creative components and outcomes through RCTs at best and A/B tests at minimum. The paper concludes with two worked examples of this process for advertisers.

1. Introduction

Creative is at the core of any advertising campaign, yet it might also be perceived as the most difficult aspect to measure (El-Murad and West, 2004). There seems to be a push-and-pull within the advertising industry, which increasingly demands more broadly encompassing approaches to campaign measurement, and yet seems reluctant to quantify the impact of campaign content. One reason for this is the difficulty of ascribing metrics to what is, at its core, a subjective experience. However, with access to more complete sources of data, the industry may be able to begin more rigorously tackling this challenge.

Summarizing Facebook Inc.'s experience understanding data to answer creative questions, the current paper proposes a framework that draws inspiration from the knowledge management literature. This framework allows the integration of knowledge from research studies as well as tacit knowledge acquired by domain experts, recognizing that creative measurement approaches need to be flexible. It explores how observations from advertising data sets can help isolate the impact of a campaign's creative content from the collection of factors that contribute to its outcome. It further examines how more complete data sources can help systematically decompose a creative asset into its component parts. Finally, two case studies underscore the importance of applying and replicating experimental methods to uncover causal relationships between creative strategies and advertiser outcomes.

2. Literature Review

Research on advertising creative content is generally performed by marketing research agencies on behalf of advertising clients. It is rare that findings from these studies are aggregated or shared, meaning few generalizable findings are available. In addition, Tellis (2009) notes that many of these studies are lab-based, with few in-market or experimental studies having been shared with the industry. Thus, methodologies used to both assess advertising content and link these features to advertiser outcome measures have been widely divergent (Naccarato & Neuendorf, 1998).

CAPTURING CREATIVE

While creative is thought to be an important factor in advertising success, there is no clear industry consensus on how to quantify this impact or how to link impact to creative recommendations. Much of this ambiguity is a product of the breadth of what "advertising creative" can refer to. It can encapsulate each and every conceptual, visual, audial, linguistic, and formal feature of a campaign, the scope of which are increasingly broad in the context of new and emerging media. There are a number of theoretical and experimental approaches that researchers have taken to build an understanding of both the impact of creative as a whole, as well as the individual features of which it is composed.

THE CREATIVE IDEA

The advertising industry generally requires that the central idea or concept behind a campaign be 'creative', and some researchers have focused on defining creativity in this context. Koslow, Sasser, and Riordan (2003) discuss the various factors that have been proposed in describing advertising creativity, concluding that the industry consensus has tended towards accepting an "originality/appropriateness"

framework. This framework seems to be a meaningful way of rationalizing the apparent trade-off between the unique and the effective.

In an effort to measure how creative an advertising campaign is, researchers have used a variety of approaches including psychometric testing, expert opinion, and advertising popularity among consumers (El-Murad and West, 2004), as well as how awarded a particular campaign is. However, it is noted in the literature that there is a marked lack of agreement in these assessments between groups. Caroff and Besançon (2008) explore the mismatch in what constitutes creativity among advertising industry professionals while Koslow, Sasser, and Riordan (2003) show that creative professionals over-weigh originality compared to consumers. Further, as noted by Peter Field Consulting[1] and System 1 consulting[2], there appears to be little consistency in the relationship between how awarded a campaign is and the advertising outcomes attributed to it.

IDENTIFYING CREATIVE COMPONENTS

Perhaps more robust measurement approaches are those used to understand the impact of specific creative components. Here, studies conceptualize the impact of creative as the aggregate impact of a number of factors on which any one creative may vary. Measurement approaches range from using principal component analysis (PCA) to reduce a large battery of messaging features to a set of creative factors (Hartnett et al., 2016), to measuring the impact of specific messaging strategies hypothesized to be important by the researcher (Bertrand et al., 2008), or measuring the impact of more objectively classifiable format features, including animation and asset size (Bruce, Murthi, and Rao, 2017).

Many of these are lab-based studies and have used content analysis and survey-based methods for classifying and quantifying creative components. In studies that use a content analysis method, each creative is coded across the same set of variables. Coding is mostly performed by researchers (Laksey et al., 1995) or subject matter experts (Haley, Staffaroni, and Fox, 1994). Other studies use survey-based classification methods. Here, researchers sample an available consumer population on their subjective assessment of the creative content (Ansari and Riasi, 2016). In more recent studies, researchers are beginning to use a combination of consumer survey responses and machine learning algorithms (Lee, Hosanagar, and Nair, 2018), in an attempt to approximate the subjective experience associated with creative content at a broader scale.

Other researchers have used experimental methodologies within the context of field studies to explore the impact of specific creative components. In these examples, specific creative variables are selected for study by the researcher. Selection of these variables may reference a theoretical understanding of how advertising operates, for example, exploring informational vs. persuasive appeals in messaging (Bertrand et al., 2008; Mortimer, 2008; Tsai and Honka, 2018)

Kolbe and Burnett (2001) note the potential for issues in achieving reliable measures in content analysis research where a small number of judges (generally two or three) are responsible for coding each creative. Additionally, a limitation in the approaches that use researcher or subject matter expert coders is the apparent lack of alignment between the judges making the subjective assessment (primarily with regard to messaging strategy) and the end consumer, whose subjective assessment may be more causally linked to the advertising outcome. These more traditional studies may not have the scale to capture complex relationships between the subjective responses to messaging strategy and advertising effectiveness. There may be interactions between creative variables, product category (Laksey et al., 1995), campaign targeting (Bruce, Murthi, and Rao, 2017), and an individual's previous advertising exposure (Braun and Moe, 2013) that are not accounted for because they are not included as variables in lab-based studies, or the statistical power is not sufficient to capture their impact. Further, whilst there may be an aggregate effect of creative components on advertising effectiveness, it is difficult to generalize which specific types of content might most consistently impact advertising outcomes (Bertrand et al., 2008; Hartnett et al., 2016), as well as the holistic impact of advertising creative on advertising outcomes.

DIGITAL VS. TV CREATIVE

Because of its dominance in modern advertising over the previous half a century, TV is the main medium for most of the creative research conducted to date (Bruce, Murthi, and Rao, 2017). Most of this research establishes the critical role of creative executional elements in affecting TV outcomes. While we expect that some of these learnings translate to digital, we also recognize that the two platforms are different enough to warrant diverging approaches to creative research on and for digital. Because of the format limitations of TV

1 https://ipa.co.uk/media/7699/ipa_crisis_in_creative_effectiveness_2019.pdf

2 https://system1group.com/blog/testing-in-the-lions-den-the-toplines

advertising, much of the work on the impact of creative concerns messaging strategy, rather than format features, or the interplay of creative and audience targeting, which can be more varied in digital advertising.

There is also some evidence that the impact of creative executional variables on outcomes for TV ads are not always generalizable across other formats (Snyder and Garcia-Garcia, 2016). Additionally, when measuring the impact of digital advertising, more care must be taken to account for potential correlation between an audience's online behavior and exposure to both the campaign at large, and a range of specific creative executions within a campaign, as the pattern of causation may not be unidirectional (Braun and Moe, 2013).

CREATIVE ATTRIBUTION

A number of researchers have used a combination of methodologies to represent the contribution of creative (or specific creative components) to campaign outcomes. However, the variation in measures and methods used, creative variables isolated, and other campaign variables controlled for make it difficult to compare studies and to make generalizations across a broader set of campaigns (Laskey, Fox, and Crask, 1995).

For example, van den Putte (2009) found some evidence for an effect of messaging strategy on brand outcomes (campaign recall and appreciation) over and above media spend. However, they note that previous purchase behavior explains a large proportion of variance in their regression models, and that media spend is correlated with market share. Further, their models that showed a high impact of messaging content strategy accounted for less than one third of the total variation in the outcome measure, suggesting scope for the inclusion of additional creative or campaign variables that may better explain these outcomes. In an alternate finding, using survey-based data, Ansari and Riasi (2016) found that advertising message and creativity was the second most impactful factor affecting brand advertising effectiveness, after media selection.

Braun and Moe (2013) examined the differential effects of a range of creative executions on online advertisement outcomes (web visits and conversions), accounting for an individual user's previous creative exposure (within the campaign). However, this study included fifteen unique creative executions, and did not allow the researchers to observe the impact of any specific creative variables on outcomes. While the study did see variation between the individual creatives' performance, it was not able to account for a collectively exhaustive range of creative variants, and therefore does not present a holistic perspective on the impact of creative. In a study that sought to account for the interplay of creative and audience targeting, Bruce, Murthi, and Rao (2017) use a dynamic model to estimate the impact of different creative formats and strategies on digital advertising engagement. As in the previous study, there was no attempt to exhaustively account for potential creative variation, and therefore it cannot provide a perspective on the overall potential impact of creative on business outcomes.

A whitepaper from Nielsen Catalina[3], often referenced in industry material, attempts to more completely account for the impact of creative in the context of the other factors that might impact advertising outcomes. This study found that of sales driven by advertising, 47% could be attributed to creative, the largest of all the advertising elements included in the study (the others being advertising reach, brand, recency, targeting, and context). It is not completely clear how the study measured "creative quality", though the paper does refer to a sales productivity metric as one of the major factors used in the analysis.

Bertrand et al. (2008) observe that many of the findings around the contribution of creative to campaign outcomes arise from lab-based experimental studies, while many of the findings approximating the impact of other factors (including media spend, allocation, and targeting) arise from observational field experiments. The following paper aims to contribute to the field by combining the two methodologies to provide a framework for hypothesis generation that is grounded in observations of the impact of creative components on an outcome measure that can be attributed to more holistic variation in advertising creative.

3. Creative Testing Framework

Many advertising and marketing organizations have different ways of determining best practice for campaign creative. Often, however, intuitive ideas that have been associated with positive results but have not been proven to be causally related to these results are elevated to best practice status (O'Dell and Grayson, 1998). Here, we propose drawing on the knowledge management literature, which outlines a process for systematically defining "best practice" within an organisation or an industry (see Figure 3.0.1 below). In fact, it seems the advertising industry is one of several uniquely placed

3 https://www.ncsolutions.com/wp-content/uploads/2017/09/NCS_Five-Keys-to-Advertising-Effectiveness.pdf

to benefit from such a process, as it consists of project-based organizational forms, which allow for low-cost experimentation and are strongly reliant on knowledge-sharing across teams (Sydow, Lindkvist, and DeFillippi, 2016).

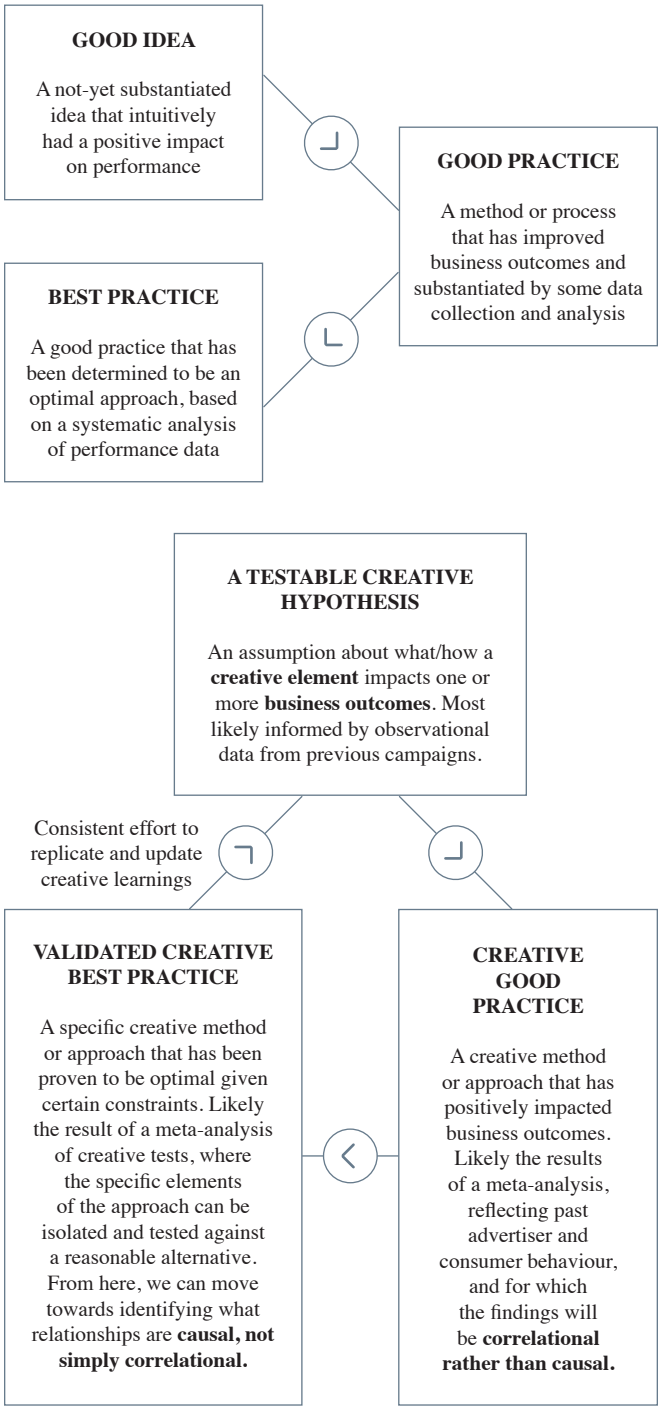

Figure 3.0.1: A frequently referenced process for defining best practice, from O'Dell and Grayson (1998), and an application of this process to the challenge of identifying creative best practice in the advertising industry.

Our recommendation is that a creative testing framework be outlined into three high-level concepts that aim to systematically validate these good ideas for creative:

- Creative Attribution: Isolate the impact of the creative on business outcomes
- Creative Components: Break down the creative to its component parts
- Experimentation: Run experiments to prove causality

Note that domain knowledge from creative professionals is important throughout the measurement framework and should inform each stage.

3.1 Creative Attribution

Before running a campaign, advertisers make certain key decisions. They decide which creative to include in their campaign, whether to target a broad or narrow audience, what objective to optimize for, where to place the ad, how frequently to show the advertisement to a consumer, and a variety of other settings. After the campaign ends, the question of which of these settings drove business outcomes and by how much is difficult to answer. For example, when exploring creative impact, the question may be, what portion of a six-point lift in an outcome is due to the creative, as opposed to the media placement? It is often difficult to say definitively what impact the creative has, unless we isolate it from the myriad of other campaign attributes.

CREATIVE EXPERIENCE

One of the ways in which the Facebook auction system optimizes for synergies between advertiser and consumer is by taking a measure of consumer value into account when deciding which advertisement to deliver to an individual user. Consumer value helps improve the user experience by identifying advertising content that people will find more relevant and interesting.

And so, as we approach the creative attribution problem, defined as isolating the impact of creative on business outcomes, it is necessary to pivot from discussions about objective "creative quality" to the more subjective "creative experience", in order to understand the interplay between an advertisement and the user who sees it. As noted by El-Murad and West (1994), capturing subjectivity is one of the challenges of creative measurement and previous research has identified the target audience as an appropriate judge of advertising creative content.

Creative Experience, in this context, is defined as how well an advertisement fits into a user's current activity session, based on a combination of polling users in the ad's target audience and other machine learning signals. This proxy metric is how we ensure that the user experience on the platform is not degraded by low quality content.

Methodology

To understand the impact of this new measure of creative experience on outcomes, we first need to predict outcomes. Using historical brand lift experiments that have run on the Facebook platform, we built a predictive model to estimate the Ad Recall Lift of an ad, as measured by Facebook's Brand Lift product[4]. Facebook's lift products are an implementation of randomized controlled trials (RCTs) tailored to the digital ad auction environment that allows advertisers to understand the incremental effect of their Facebook advertising.

For Ad Recall Lift[5], the question of interest that is asked of and compared between the test and control groups is whether the user "recalls seeing an advertisement for [Brand X] online or on a mobile device in the last 2 days?" Ad Recall lift is designed to measure how much more likely the user is to recall the brand's advertisement after having an opportunity to see it and has been proven to be correlated with other lower funnel metrics such as brand favorability, purchase intent and, in some cases, sales. Foundational research by industry measurement leader Nielsen[6] has shown that attitudinal brand metrics can be predictive of sales. It is also important to note that while most advertising metrics are measured at the campaign or study level, this methodology uses modeling to decompose campaign effects to the level of the creative asset (rather than the campaign or set of campaigns as a whole). This allows us to isolate the creative impact not just for a campaign but for each asset within it. Additional campaign features the model controls for include the advertiser vertical (product category), the campaign optimization (the result the advertiser chose to prioritize in the ad delivery auction), the placement (which surface the advertisement appeared on, for example Facebook, Instagram, or Messenger), the creative format (video or static image), the type of targeting (either a general age, gender, or location target or more specific interest-based targeting), and broad geographical region. These metrics are all encoded by the advertiser or the Facebook advertising system during the campaign set-up process.

Model

After investigating several other methods for predicting ad recall lift, the best performing model was a Gradient Boosted Decision Tree (GBDT). GBDT is a boosted ensemble decision tree algorithm that minimizes the loss function using gradient descent (Friedman, 2001; Friedman, 2002). GBDT allows us to capture the non-linearity of our outcome variable and the interaction effects between features. The model is trained with a 70/30 train/test split using Facebook Brand Lift studies from an 84-week look-back window. The model is updated weekly.

Inverse probability weighting is used to reduce the campaign attribute skewness. The proportion of examples in the training dataset with the same combination of features (vertical, optimization, placement, creative format, targeting and region) is calculated and then the inverse of that value is applied to each observation to weight it.

The model is trained to predict *Ad Recall Lift using the following features: 3 second View Rate, Placement, Creative Format, Optimization, Targeting Type, Business Region, Vertical, Creative Experience Score*.

Counterfactual Simulation

Using our trained model, we can apply counterfactual simulation to isolate the creative impact. Counterfactual simulation measures what happens to our predicted outcome variable when we simulate a change in one of its dependent features. To do this, we start with one row of data and generate synthetic rows of data from it, where the rows are exactly alike except for the feature of interest. We then run these rows of data through our trained model to see how the predictions for one row might differ from another. An example of this synthetic data for an asset can be seen in Table 3.1.1.

4 https://www.facebook.com/business/help/1693381447650068
5 https://www.facebook.com/business/help/310485426154135
6 https://www.nielsen.com/us/en/insights/article/2011/research-shows-link-between-online-brand-metrics-and-offline-sales/

DATA	AD	ATTRI-BUTES OF AD	CREATIVE EXPERIENCE OF AD	PREDICT-ED LIFT OF AD
Synthetic	Ad XX	Same	0	1.2
Synthetic	Ad XX	Same	1	1.8
Synthetic	Ad XX	Same	2	1.83
Synthetic	Ad XX	Same	...	...
Original	Ad XX	Same	37	2.4
Synthetic	Ad XX	Same	...	...
Synthetic	Ad XX	Same	100	3.9

Table 3.1.1.: An example of synthetic data created for a hypothetical creative asset with Creative Experience = 37 and Predicted Ad Recall Lift = 2.4.

Using the trained model for our example, we run predictions for each ad, with every possible value of the creative experience score [ranging from 0 to 100]. This allows us to establish upper and lower bounds on how much the ad recall lift changes with creative experience, i.e. a Creative Lift Potential.

$$\textit{Creative Lift Potential}_{Ad\ XX} = \textit{Max (Predicted Lift}_{Ad\ XX}\textit{) - Min (Predicted Lift}_{Ad\ XX}\textit{)}$$

Note that we could not do this simply by using predictions for the lowest (0) and highest (100) values of Creative Experience because our model is non-linear and so Predicted Lift does not necessarily linearly increase or decrease with an increase in Creative Experience.

Additionally, because the predictions are made at the level of an individual ad, the Creative Lift potential will vary for each creative asset based on all of its other features, e.g., Vertical, Optimization, and Placement. We can then compare the Creative Lift Potential to the prediction for the original asset to understand how much more the creative can be improved to drive lift. This process is illustrated in Figure 3.1.2.

Creative Lift Potential =
Lift Currently Driven by Creative + Additional Lift That Can Be Driven by Creative

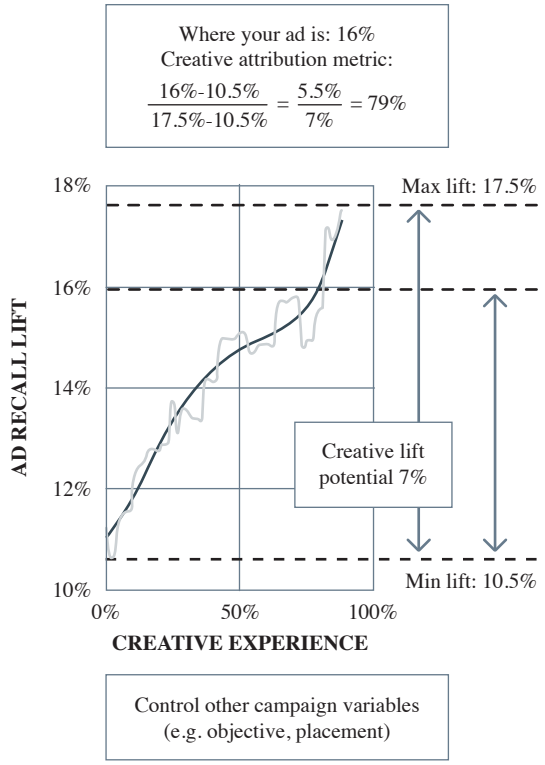

Figure 3.1.2.: The calculation used for determining the creative lift potential for a hypothetical creative asset.

Validation

We validated the accuracy of the Lift prediction model by comparing predicted to actual values for the 30% holdout set. Results can be seen in table 3.1.3. We see a high correlation between our predicted and actual values of 0.66. We also see 73% accuracy in lift predictions, that is how often the model is able to correctly predict the creative asset that performs best in the campaign. In error margins, we see a median error of 24% and median absolute error of 53%, both of which measure how close our predicted lift comes to observed/actual lift and both of which compared favorably to other internal efforts to model brand lift. These metrics measure bias in our model and tell us that our model predictions consistently slightly overestimate lift. Also, even though the sample graph above in Fig. 3.1.2 appears non-monotonic and overfit to our data, the smoothed predicted line is generally increasing and within the confidence intervals of our prediction function.

ATTRIBUTION SYSTEM	MODELED BRAND LIFT
Winner Accuracy	72.9%
Median Error	23.6%
Median Absolute Error	52.8%
Estimate within 30pct	32.2%
Correlation	.66

Table 3.1.3: Validation metrics for creative attribution model.

Application

This modeled creative attribution metric is one such way to isolate the impact of creative on outcomes to measure creative impact across campaigns more broadly. However, as we show in the case studies below, creative features can also be tied to outcomes directly using simpler regression-style models, to answer specific creative hypotheses.

3.2 Creative Components

As we mentioned earlier, directly quantifying the "creative quality" of an advertisement is highly subjective. While many people may be capable of describing what they find aesthetically pleasing about creative content, the full feature space that can explain this description is broad and often unwieldy. In addition, while we may be able to list these features, some are more quantifiable than others. Concepts such as color composition, aspect ratio, and the presence of people or animals may be something we can easily distinguish, whereas other creative elements such as emotion and narrative can be difficult to describe and can differ between individual viewers. Additionally, once we start to scale up the collection of features, we may learn that teaching an automated system to detect certain features can be extremely difficult.

To gain a better understanding of the scope of the creative feature space in advertising, and how we may think of collecting them, we have grouped creative elements into three different categories.

- **Mechanical:** Static and video advertisements have creative elements that can be described by objective numeric metrics. Creative elements such as the aspect ratio of the asset, potential play time of a video, average color, and resolution are descriptions of an asset that have one true value.
- **Visual:** The presence of physical objects such as people, animals, or cars, or if the setting is indoors, outdoors, or synthetic, or the use of overlaid text as an image are concepts that describe the visual components of an ad. These may not be as straightforward to quantify as mechanical metrics, but they still have objective answers.
- **Thematic:** The highest-level description of a creative campaign. Creative assets can be designed to evoke a particular emotion from the viewer. They can also be designed with different production choices in mind: smooth or choppy, fast paced or slow and methodical. However, the emotions one person feels when viewing a creative asset may be different to someone else's. This results in many of these features being highly subjective; many thematic features depend on the context they are being viewed in and who is viewing them.

A large majority of the creative aspects of an advertisement can be categorized into one of these groups. While we talk about both static and video assets in the same way, there are some differences in how the creative features are measured for both. For static assets, the features are collected on the single image. For video assets, the features need to account for the multiple frames found in a video. Each frame can be viewed as a static image resulting in a large number of measurements depending on the length of the video. However, since we generally want to represent each video as a single set of creative features (i.e., one measurement for the presence of people in an ad), we can think of averaging the creative features over every frame of a video, or simply averaging over "key frames" that we detect in the ad.

The categories of creative features we have described are how we will think about describing creative assets, but they can just as easily be used in a non-advertising setting to describe general photos or videos. In fact, virtually all of the technology that is used to identify components such as these were created in advertising agnostic settings. In addition, this breakdown of features is high-level, and we don't fully discuss the fact that depending on the platform, you may have some creative aspects of your advertisement that are specific to where you are advertising. This can include aspects such as the context the advertisement appears in, the addition of your logo or company name, and copy or titles that are separated from the static or video creative. While we don't discuss these aspects here, it is possible to use details about the advertising platform to determine how these features could be categorized into the above groupings.

Mechanical, visual, and thematic creative components are ordered by the complexity of identifying them in creative assets. Mechanical components are typically the easiest to identify and can be described by virtually every off-the-shelf piece of design software and many libraries or built-in functions in various programming languages. Visual and thematic elements are more difficult to measure as they often require sophisticated machine learning algorithms implementing computer vision methods to detect different visual or thematic elements. The major differentiator between these two categories is the type of data that the underlying machine learning algorithm will require. In both cases the algorithms we use will require a training set of image or video data where we know whether or not a given visual or thematic element is present. For visual elements this is slightly more straightforward as the subjects we are defining (e.g., people, animals) are more objective and therefore easier to describe. A major limitation here is collecting a sufficient set of labeled data points, although for many visual concepts training data sets are widely available. On the other hand, thematic elements pose a much larger challenge as they are more subjective. For example, an advertisement that is humorous can be humorous in many different ways and collecting a ground truth dataset that fully describes these possibilities can be very difficult.

The set of potential creative features that can be used to describe a creative asset is very broad. But, advertisers looking to describe creative quality and its impact on their business must determine which features are practical. One method to reduce the size of the feature set is to start to narrow down a potential set of useful creative features by first determining which features we can reliably measure and then which features we can reliably action. However, not all advertisers will know how or what can be reliably measured or what creative features can be manipulated due to different technical or organizational roadblocks. To accurately assess this requires a level of domain knowledge of measurement and creative design.

APPLYING DOMAIN KNOWLEDGE

More formally, we have defined four criteria for the selection process of creative elements that should be experimented on: (1) they can be meaningfully adjusted, (2) they can be identified as important by creative professionals, (3) they can be measured and quantified, and (4) they can be recommended to advertisers. These criteria help ensure that creative elements are selected that will allow advertisers to take action and explore the impact of creative. For example, the mechanical creative feature of video asset duration is often seen by creative professionals as a key creative aspect when designing video assets. As this feature is mechanical, it is easy to measure. It requires some creative expertise to manipulate and verify that a creative asset still communicates a meaningful message at a different video length. Depending on the specific advertiser, features such as video length, or other mechanical features, may often be selected by creative professionals with domain knowledge as they pass the four criteria we have mentioned. More complex elements like brand-specific characters, emotion, or use of humor, often require higher production costs and might necessitate reshooting a campaign. This may be prohibitively resource- intensive for most advertisers.

3.3 Experimentation

The key questions we are attempting to answer when exploring the creative quality of an advertisement are "does the design of my creative drive more value for my business" and "what features of my creative asset should I change to maximize positive impact?" To make effective decisions about the design of a creative, where we attempt to maximize the potential impact it can have on business outcomes, randomized experiments should be used. Experimental setups, which have been fine-tuned in fields such as medicine, epidemiology, and psychology, have a long history of being the most efficient method to measure the potential impact of new decisions (Levitt and List, 2009; Gordon et al., 2019). This is due to their ability to isolate and measure the effect of individual design choices.

Consider an example where we are testing the impact of an individual creative asset on sales for a specific company. The most precise way of measuring this effect is to randomly split the target audience for the campaign into two groups: A and A'. Group A will get served the asset we designed during the campaign and group A' will not be served the ad. Since these groups were made at random, they are equivalent and comparable across features such as age, gender, and occupation—the only difference in the two groups is that one group received the asset being studied and the other did not. After the campaign ends, if we compare the sales rate between both groups, we can be reasonably confident that any observed difference is due to exposure to the creative asset since all other features are equivalent, on average, across groups.

This type of setup, where randomization is used to build comparable groups, is important for answering questions around effectiveness. If we were to simply serve advertisements to our target audience but not maintain a comparison group of people at random, we would have no equivalent group against which to compare outcomes such as sales rates. The users targeted to receive the advertisement were most likely targeted since they are already possible consumers of the product for sale; they will be more likely to buy the product even without an ad. People who were not exposed are generally not part of the target audience. Comparing the sales rates between these groups would be an unfair comparison as attributes such as age or gender could be the actual reason for any differences.

This experimental setup can be extended to compare two individual creative assets. Consider another example where we want to test whether having a company's logo placed in the bottom right corner of an image leads to more sales. In an experimental setup, we design two creatives that are identical in every way except that one has a logo in the bottom right corner, and one does not. We have a few options for how to conduct this type of experiment, but one method would be to split the target audience into four groups at random: Groups A and A' where users are exposed to the logo asset or no advertisement at all, and groups B and B' where users are exposed to the non-logo asset or no advertisement at all. This setup allows us to measure the impact of the logo asset (by comparing groups A and A') as well as the impact of the non-logo asset (by comparing groups B and B'). The results of a study like this can give us information about the effect of logos in a company's ad.

When setting up experiments, there are a few design choices that will result in different types of measurement. In the setup we just described to test logos, the experiment creates two treatments: one with a logo and one without. We also setup two control groups, to whom no advertisements are served. This variant on a randomized control trial gives us a measurement of how effective each treatment was in isolation by making comparisons to no advertisement control groups. An alternative to this setup would be to create only two groups: one for the logo asset and one for the no-logo asset. In this A/B test setup, we do not have a control group of no advertisements and we instead directly compare the logo asset group to the no logo asset group. The comparison here is different as both groups are exposed to some creative asset, the only difference being the presence of a logo. Both of these experimental setups are popular and have various benefits and drawbacks that we don't cover here. For a more in-depth exploration into experimental design, see Gordon et al. (2019)

If we use an experimental setup as our framework for testing the effect of creative decisions, how do we then determine which creative decisions to test? As mentioned earlier, the potential size of the creative feature space is vast, but we can narrow down creative features by first determining what is measurable and actionable. However, we may still be left with a large set of potential features to test. One way to further narrow down the creative feature space is to use correlational meta-analyses or observational studies to determine if there are creative features that have some relationship with a business outcome of interest.

Given a collection of experiments run on different campaigns, we could take a high-level view and see if any of our identifiable and actionable creative features are correlated with good experimental outcomes. For example, we could use a linear regression to determine the relationship between creative features and the results of a sales experiment. While these relationships are only correlational, if we apply domain knowledge of the advertising platform we are working with, we can explore the full set of creative features and find those that have strong relationships and that are reasonable to design experimental tests around.

This type of approach is popular, but also suffers from being an aggregate view of the importance of different features as they are all lumped into one model. However, in recent years, there has been a push to make the relationships learned by models, such as regressions, more interpretable at an individual level using methods such as local interpretable model-agnostic explanations (LIME) (Ribeiro, Singh, and Guestrin, 2016) or Shapley values (Lundberg and Lee, 2017). In our advertising examples, we can explore the most important creative features for individual assets. This additional granularity allows us to fine tune our correlational hypothesis before designing and conducting experiments. We could take all the creative assets for an individual advertiser or for a target audience and learn important creative features for them. This way we can run more efficient experiments that have clear, generalizable implications.

APPLYING DOMAIN KNOWLEDGE

One of the critical components of running a creative experiment is ensuring comparability of the assets between cells and ensuring that the variable of interest

has been sufficiently isolated. This often involves creative professionals designing a baseline version of a creative asset that makes it easy to manipulate the element of interest for each cell. This ensures that the test is not confounded by other elements within the creative, to the extent possible. However, even when we isolate a single creative component through these methods, we might still see outcomes that are surprising or unclear because creative can have so many small variations. This often requires a creative professional to draw up additional hypotheses to address alternate explanations for the results we see. These hypotheses can help us design follow-up tests to tease out confounding factors or understand more about the dynamics of the effect of a creative element.

4. Case Studies

The following two case studies demonstrate how the creative testing process can operate in practice, as executed by a number of cross-functional teams. We explain the application and the impact of the process for identifying two important creative considerations for the Facebook advertising platform.

4.1 Focal Point

OUTCOME

The creative attribution signal described in section 3.1. provides a useful outcome metric for understanding the potential impact of a set of creative features at the level of the creative asset. We can then observe the variation in the outcome with reference to creative components of each individual creative asset.

CREATIVE COMPONENTS

Creative rating method

This study used a creative rating method whereby advertisers elected to apply measures from a set of eight creative features to their own campaign assets. Each creative asset submitted was coded on a set of seven metrics by Amazon Mechanical Turk[7] raters. Validation work showed that forty raters was about the point where results stabilized, therefore forty raters scored the creative on each variable on a five-point Likert scale within the context of a mobile Facebook feed environment. The final score used for each metric is the number of raters who registered agreement with the variable (points 1 or 2 on the scale). The metrics used in the analysis are defined in Table 4.1.1.

CATEGORY	FEATURE	ITEM
VISUAL	Noticeability	This ad would grab your attention
	Focal Point	This ad has one obvious focal point
BRANDING	Brand Association	It is easy to identify the advertiser in the ad
	Brand Fit	The ad fits with what you know about the brand
MESSAGING	Message Comprehension	It is easy to understand the message
	Emotional Reward	This ad appeals to you emotionally
	Call-to-Action	This ad urges you to take a clear action
	Interesting Information	This ad has interesting information

Table 4.1.1: A description of the creative metrics on which each asset was rated.

EXPERIMENTATION

Meta-Analysis

To understand which creative features had a potential relationship with the outcome metric, a meta-analysis was conducted. The analysis included 3,000 static creative assets which were active for between 3 and 90 days on Facebook feed and for which the advertiser had collected creative diagnostic metrics. To identify any relationships between each creative metric and the outcome variable at an aggregate level, a weighted least squares regression model was used. The model incorporated the overall lift driven as the dependent variable, controlling for lift potential and campaign characteristics:

$$Lift\_driven = \beta 0 + \beta 1\ Lift\_potential + \beta 2\ Targeting + \beta 3\ Vertical + \beta 4\ Region + \beta 5\ Bid\_type + \beta 6\ Advertiser\_type + \beta 7\ Campaign\_duration + \beta 8\ Year + \beta 9\ Avg\_CPM + \beta 10\ Objective\_type + \beta c\ Creative\_elements + \varepsilon$$

Results of the regression analysis for the variable of interest can be found in Table 4.1.2. Additional campaign characteristics that were controlled for included the type of targeting (either a general age, gender, or location target or more specific interest-based targeting), the advertiser vertical (product category), broad geographical region, the number of days the campaign was in market, the average spend per 1,000 impressions, the year of the campaign, bid type (how the advertiser specifies their bid in the Facebook advertising system), whether the advertiser was a very large organization or not, and

7 https://www.mturk.com/

finally, the campaign objective (a direct response or brand objective). These metrics are all encoded by the advertiser or the Facebook advertising system during the campaign set-up process. The regression model explained ~50% of the variation in lift driven (by the creative experience). "Emotional reward", "focal point", and "noticeability" were significantly positively correlated with the outcome variable (at p<0.001).

MODEL COEFFICIENTS
DEPENDENT VARIABLE = LIFT DRIVEN

	Estimate	Significance level
Lift potential	0.32	***
Creative Element: Brand association	-0.03	***
Creative Element: Brand fit	0.01	
Creative Element: Call to action	-0.01	
Creative Element: Emotional reward	0.03	***
Creative Element: Focal point	0.02	**
Creative Element: Interesting Information	-0.03	***
Creative Element: Noticeability	0.04	***
Model Adjusted R^2	0.50	***

Significance codes: 000 ***; 0.001 **; 0.01 *; 0.05 .

TABLE 4.1.2: Regression output for creative element meta-analysis

A cross-functional business team, including creative strategists and measurement experts, worked to understand the results of the analysis and design a testing plan to determine causal relationships between creative variables of interest. In the regression analysis, the "focal point" element was significantly correlated with outcomes, yet looking at examples of creative assets that had high "focal point" scores, creative strategists were unsure as to whether this was due to the presence of a singular focal point, or because in most of the images, the product was front-and-center. Further, focal point was identified by the creative strategists as a variable that could be manipulated and observed across a broad spectrum of brands and campaigns.

The resulting research question was framed as:

Does creative with a single, visual focal point (vs. product-focused creative) drive greater advertising outcomes (brand awareness or direct response)?

RANDOMIZED CONTROL TRIALS

Based on the aforementioned meta-analysis and domain knowledge, it was hypothesized that:

1. There is a positive relationship between single focal point and brand awareness lift (controlling for high/low product use)
2. There is a positive relationship between single focal point and view content events (controlling for high/low product use)

Study Design

A 2 (single focal point; many focal points) X 2 (high product use; low product use) between-subjects design was used. Each experimental cell included one creative asset (see Figure 4.1.3) for the same product, with the same messaging across all cells. All other campaign features were consistent across all four cells. For each campaign, both a Facebook Brand Lift and a Facebook Conversion Lift study were run to measure incremental brand awareness and incremental view content events.

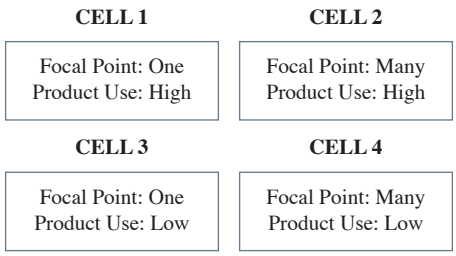

Figure 4.1.3: Experimental cells for the Focal Point experiment

Sample

Sampling for brand lift and conversion lift tests was achieved in line with standard usage of these products (see section 3.4). The study was replicated across 14 Ecommerce advertisers who were invited to participate.

Creative assets

For each advertiser, two images were selected, one product-focused image and one lifestyle image. Images were selected to be similar in style but different in content. Focal point was manipulated by cropping or blurring the image to focus on a single visual feature. To test whether the focal point dimension had been successfully manipulated, each creative asset was evaluated within the creative diagnostic tool. The focal point score was significantly higher for the single focal point assets (p<=0.05).

Results

To evaluate the results across the set of 14 advertisers, the creative manipulations were compared on how many

times they ranked 1st or 2nd based on the outcome measures. Because each study was small, both 1st and 2nd place were identified as "winners" for the purpose of the meta-analysis. One study was discounted from the analysis as spend levels differed significantly across cells. Of the remaining 13 studies, there were 52 total cells which could have been identified as a 'winner'.

For the brand awareness outcome, 62% of high focal point creatives were winning creatives compared to 38% of low focal point creatives. A randomized block ANOVA with two independent variables (focal point and product use) showed a positive relationship between focal point and brand outcome which was significant at $p<0.1$. High focal point creatives were 60% more likely to be 'winning' creatives on the cost-per-brand outcome metric.

For the conversion outcome, 87% of high product use creatives were winning creatives compared to 27% of low product use creatives. A randomized block ANOVA with two independent variables (focal point and product use) showed a positive relationship between product use and brand outcome which was significant at $p<0.1$. High focal point creatives were 87% more likely to be 'winning' creatives on the cost-per-conversion outcome metric.

From this study, we see evidence for a positive, causal relationship between brand awareness and high focal point within a creative asset. We also see evidence for a positive, causal relationship between view content events and high product use.

Implications

This study suggests that, on average, advertisers looking to achieve brand outcomes can employ the use of a clear, single focal point to improve their creative performance. Alternatively, if the advertiser is looking to achieve lower funnel, or direct response outcomes, clearly featuring product may improve outcomes over and above featuring lifestyle imagery. As these results were related to studies among 13 Ecommerce advertisers in the US market, additional experimentation would be needed to understand whether the results generalize across geographic markets and advertiser product categories. Further, we expect creative preferences to change due to factors including time, competitors, market growth, visual and creative trends. Therefore, as the context of these creative executions shift, we expect these results may also evolve.

4.2. Mixed Format

OUTCOME

Our outcome of interest in this study is brand lift, as we are looking to understand the ideal composition of assets within a campaign and so a campaign-level metric is needed.

CREATIVE COMPONENTS

Our creative feature here is the creative format of an advertisement and how a mix of multiple formats in the campaign can affect outcomes. Creative formats include static, video, carousel, and canvas. More information, including format descriptions and examples can be accessed on the Facebook for Business website[8].

DOMAIN EXPERTISE

Setting up this experiment was particularly challenging, because it attempts to compare across creative formats. As such, it was vital that the creative assets were designed such that the longer formats do not encode any more information that might make them perform better, outside of the fact that they are longer. Our internal Creative Shop team was vital in ensuring comparability across the creative assets, making sure they had the same look and feel and felt like they were part of the same campaign, whether they were video or static.

EXPERIMENTATION

Meta-Analysis

An analysis of over 3,000 brand lift studies in the US that ran between September 2017 and August 2018 was conducted to understand the effect of multiple formats in a campaign on brand lift outcomes. Based on the formats of assets included in each campaign (formats: static, video, carousel, canvas), each campaign was classified into:

- Single assets if the campaign had only one format
- Mixed assets if the campaign had multiple formats

Analysis was limited to only lift studies where results were statistically significant, and variables such as campaign duration, frequency, vertical, country, campaign budget, optimization goal, and other logged campaign features were controlled for.

8 https://www.facebook.com/business/help/1263626780415224?id=802745156580214

Higher lift across all of the upper funnel brand metrics was observed for campaigns that used a mix of creative assets, compared to those that used video only or static only. An index of Lift statistics for different mixed-asset cells compared to single-asset cells are broken down by brand outcome metrics in Table 4.2.1.

US/ALL VERTICALS/ALL PLATFORMS

Question Type	Index: Mixed-Asset Cell vs. Single-Asset Cell
Ad Recall	1.17
Top of Mind Awareness	1.17
Message Association	1.32
Familiarity	1.79
Affinity	1.92

Table 4.2.1: Index of Lift study results for single vs. mixed asset campaigns for different Brand Lift questions.

However, outcomes cannot be compared at face value for these groups of advertisers because endogenous effects likely still exist outside of the factors that were controlled for. Therefore, determining causality of these findings requires us to run experiments to validate the findings.

RANDOMIZED CONTROLLED TRIALS

Based on the aforementioned meta-analysis, it was hypothesized that: Campaigns with mixed formats will drive higher Ad Recall Lift than campaigns with a single asset type.

Study Design

A two-cell test plan (see Figure 4.2.2) was created with one cell using a single creative asset and the other using a mix of creative assets. The single creative asset campaigns had one or more advertisements using only one format - static or video assets with the same duration, but not both. The mixed creative asset campaigns had a combination of creative assets that were produced using different advertising formats or multiple videos of different lengths. Thirteen advertisers were recruited to participate in the study. For each campaign, a Facebook Brand Lift study was conducted to measure incremental lift in ad recall.

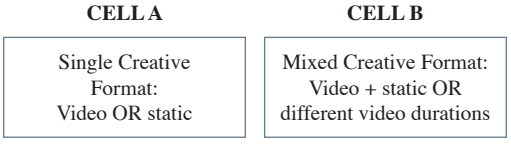

Figure 4.2.2: Experimental cells for the Mixed-Format experiment

Results

Eleven of the thirteen campaigns showed a statistically significant difference in ad recall lift between cells. The mixed creative format cell drove higher incremental ad recall lift than the single format cell in ten out of these eleven experiments. On average, for studies in which a statistically significant result was achieved, the mixed-format creatives required 52% fewer resources per incremental ad recall.

Implications

These experiments showed that the use of mixed creative assets can result in higher ad recall lift than using single formats, improving advertiser value overall. While we could hypothesize that exposing the consumer to different formats may help reinforce the campaign message, we cannot definitively say why a mix of creative assets performed better. The recommendation for advertisers is to develop creative assets of different formats (or videos of different durations), instead of single creative assets, to expose audiences to different assets within their campaigns.

However, we also recognize that this experiment was conducted for a small sample of advertisers. Validating whether these findings are generalizable across all advertisers would require replicating the experiments across a larger and more representative sample.

5. Summary & Implications

Because of the perceived ambiguity and difficulties with creative testing, most advertisers currently rely on “good practice” and intuitive knowledge and never push toward “best practices”. While there are many reasons why this is the case, the authors of this paper advocate for the use of data-driven solutions to home in on creative features that can meaningfully impact advertiser outcomes, followed by the implementation of a creative testing framework. The current paper hopes to illustrate that with the right framework it is possible to test a wide variety of creative variations on digital platforms, making creative testing viable and practical for most advertisers and encouraging them to innovate to improve their business outcomes. Additionally, the framework outlined is not executionally strict, provides flexibility in implementation and has broad applicability for digital advertisers and advertising platforms alike. At its core, the framework advocates for:

- Isolating and measuring the impact of creative separate from all of the other factors that can affect business outcomes;
- Deconstructing the creative to its component elements and testing creative elements one at a time;
- Running RCT experiments to prove causal relationships between the creative element being tested and the outcomes that result;
- Underscoring the entire process with creative professional domain expertise informing each stage.

This framework acknowledges the reliance of creative research to date on domain knowledge and proposes not its replacement, but the addition of more rigorous measurement and experimental frameworks to improve the quality of findings and understand the breadth of their applications.

LIMITATIONS

Experimentation Platforms

The authors recognize that this framework relies on Facebook's industry-leading experimentation platform (with its Brand and Conversion Lift products). Many practitioners will undoubtedly face significant limitations in implementing meta-analyses that can incorporate sufficient data sources with regard to advertising outcomes. However, there are many creative consulting companies who have developed and implemented creative analytics tools that incorporate outcomes from major digital platforms[9].

Additionally, media budgets and logistical requirements for the implementation of RCTs on digital platforms may be prohibitive for many advertisers. Therefore, while RCTs would be ideal to determine the extent of causal impact, A/B tests, which can be easily implemented within Facebook's Ads Manager tool[10], can prove useful in their absence.

Finally, while computer vision technology can help identify creative variables for experimentation, it is currently not possible for these methods to reveal anything about the more conceptual nature of creativity, which many in the creative industry regard as the ultimate goal of a creative development team. However, the current authors believe that creative professionals will continue to play a key role in the implementation of creative experimentation, and results and best practice guidance can help the industry understand the parameters they should work within when developing highly creative concepts and advertising executions.

Future Research

Isolating Creative

This paper advocates for an assessment of creative on the basis of its performance as it relates to the advertiser's business goal. Therefore, it is important to tie "creative success" to business outcomes and not awards or aesthetic measures that do not accurately capture business impact. Further, isolating and quantifying the proportion of this business impact that is exclusively attributable to creative is important for good creative assessment. In the above sections, we propose a method for undertaking this by aggregating results across historical lift experiments to predict and estimate creative impact. The methodology that we have identified is one such approach and further research is needed on alternatives and the best methods of implementation.

Creative Element Detection

Current computer vision technology has proven indispensable in detecting creative elements at scale for a high volume of assets and this is where more research is needed. Innovations in machine learning algorithms that can accomplish this creative element detection at scale with high accuracy and with more complex elements (use of humor for example) will undoubtedly redefine what creative testing looks like at scale for the industry. It will also enable this testing framework to be utilized in optimizing advertisements in real-time.

Acknowledgements

We would like to thank several researchers at Facebook who conducted projects from which we developed this framework. Liyun Chen and Fernanda Alcantara for their work on the mixed-format analysis and Elif Isikman for her work on the focal point case study. Finally, we would like to thank Audrey Burgess and Adam Berger for their help in completing the paper.

9 https://www.facebook.com/business/partner-directory/search?solution_type=creative_platform

10 https://www.facebook.com/business/help/1738164643098669?id=445653312788501

References

Ansari, Azarnoush, and Arash Riasi. "An investigation of factors affecting brand advertising success and effectiveness." International Business Research 9.4 (2016): 20-30.

Bertrand, Marianne, et al. "What's advertising content worth? Evidence from a consumer credit marketing field experiment." The quarterly journal of economics 125.1 (2010): 263-306.

Braun, Michael, and Wendy W. Moe. "Online display advertising: Modeling the effects of multiple creatives and individual impression histories." Marketing Science 32.5 (2013): 753-767.

Bruce, Norris I., B. P. S. Murthi, and Ram C. Rao. "A dynamic model for digital advertising: The effects of creative format, message content, and targeting on engagement." Journal of marketing research 54.2 (2017): 202-218.

Caroff, Xavier, and Maud Besançon. "Variability of creativity judgments." Learning and individual differences 18.4 (2008): 367-371.

El-Murad, Jaafar, and Douglas C. West. "The definition and measurement of creativity: what do we know?." Journal of Advertising research 44.2 (2004): 188-201.

Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.

Friedman, Jerome H. "Stochastic gradient boosting." Computational statistics & data analysis 38.4 (2002): 367-378.

Gordon, Brett R., et al. "A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook." Marketing Science 38.2 (2019): 193-225.

Haley, Russell I., James Staffaroni, and Arthur Fox. "The missing measures of copy testing." Journal of Advertising Research 34.3 (1994): 46-61.

Hartnett, Nicole, et al. "Creative that sells: How advertising execution affects sales." Journal of Advertising 45.1 (2016): 102-112.

Kolbe, Richard H., and Melissa S. Burnett. "Content-analysis research: An examination of applications with directives for improving research reliability and objectivity." Journal of consumer research 18.2 (1991): 243-250.

Koslow, Scott, Sheila L. Sasser, and Edward A. Riordan. "What is creative to whom and why? Perceptions in advertising agencies." Journal of advertising Research 43.1 (2003): 96-110.

Laskey, Henry A., Richard J. Fox, and Melvin R. Crask. "The relationship between advertising message strategy and television commercial effectiveness." Journal of advertising research 35.2 (1995): 31-40.

Lee, Dokyun, Kartik Hosanagar, and Harikesh S. Nair. "Advertising content and consumer engagement on social media: evidence from Facebook." Management Science 64.11 (2018): 5105-5131.

Levitt, Steven D., and John A. List. "Field experiments in economics: The past, the present, and the future." European Economic Review 53.1 (2009): 1-18.

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.

Naccarato, John L., and Kimberly A. Neuendorf. "Content analysis as a predictive methodology: Recall, readership, and evaluations of business-to-business print advertising." Journal of Advertising Research 38 (1998): 19-29.

O'Dell, Carla, and C. Jackson Grayson. "If only we knew what we know: Identification and transfer of internal best practices." California management review 40.3 (1998): 154-174.

Putte, Bas van den. "What matters most in advertising campaigns? The relative effect of media expenditure and message content strategy." International Journal of Advertising 28.4 (2009): 669-690.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." arXiv preprint arXiv:1606.05386 (2016).

Snyder, Jasper, and Manuel Garcia-Garcia. "Advertising across platforms: Conditions for multimedia campaigns: A method for determining optimal media investment and creative strategies across platforms." Journal of Advertising Research 56.4 (2016): 352-367.

Sydow, Jörg, Lars Lindkvist, and Robert DeFillippi. "Project-based organizations, embeddedness and repositories of knowledge." (2004): 1475-1489.

Tellis, Gerard J. "Generalizations about advertising effectiveness in markets." Journal of advertising research 49.2 (2009): 240-245.

Tsai, Yi-Lin, and Elisabeth Honka. "Non-Informational Advertising Informing Consumers: How Advertising Affects Consumers' Decision-Making in the US Auto Insurance Industry." Available at SSRN 3094448 (2018).