Do Viewers Care?
Understanding the impact of ad creatives on TV viewing behavior

Yannet Interian, Kaustuv, Igor Naverniouk, P. J. Opalinski, Sundar Dorai-raj, and Dan Zigmond*
Google, Inc.

Abstract

Google aggregates data, collected and anonymized by the DISH Network L.L.C., describing the precise second-by-second tuning behavior for millions of television set-top boxes, covering millions of US households, for several thousand TV ad airings every day. From this raw material, Google has developed several metrics that can be used to gauge how appealing and relevant commercials appear to be to TV viewers. While myriad factors impact tuning during ads, we find a measurable effect attributable to the ad creative itself. Although this effect appears modest, it demonstrates that viewers do react differentially to TV advertising, and that these reactions can then be used to rank creatives by their apparent relevance to the viewing audience.

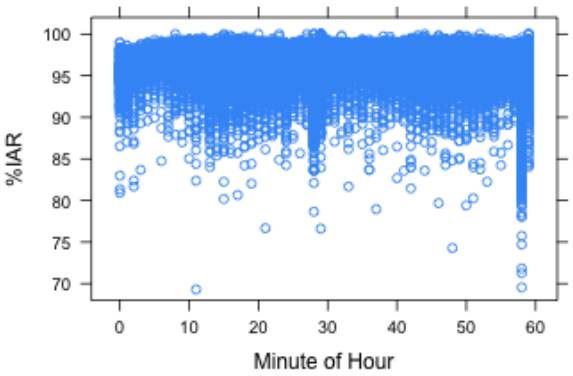

Figure 1: %IAR as a function of the minute of the hour the ad was aired.

1 Why viewers tune away

Google has developed several metrics based on second-by-second tuning data collected from several million US television set-top boxes[1]. This paper focuses on the most promising of these, the percentage of initial audience retained (%IAR) during a commercial. This is calculated by taking the percentage of the TVs tuned to an ad when it began which then remained tuned throughout the ad airing[2].

The intuition behind this metric is that when an ad does not appeal to a certain audience, viewers will vote against it by changing the channel. By including only those viewers who were present when the commercial started, we hope to exclude some who may be channel surfing. However, even these initial viewers may tune away for other reasons. For example, a viewer may be finished watching the current program on one channel and looking for something else to watch.

For example, the chart in figure 1 shows %IAR val-

*Please address correspondence to djz@google.com.

[1]These anonymous set-top box data were provided to Google under license by the DISH Network L.L.C. and Google gratefully acknowledges their assistance in making this work possible, and particularly Steve Lanning, their Vice President for Analytics, for his helpful feedback and support.

[2]We have also calculated these metrics based on households rather than televisions. The results are nearly identical because we find it is unusual for multiple TVs in the same household to be watching the same ad.

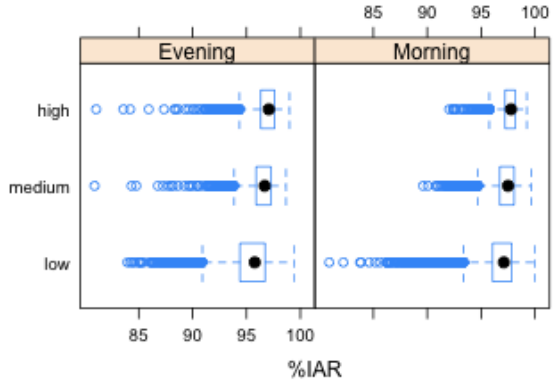

Figure 2: Impact of initial audience size on %IAR.

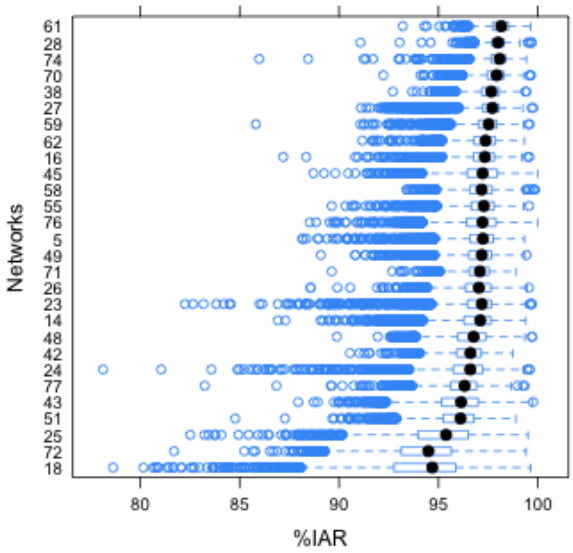

Figure 3: Impact of the underlying network on %IAR

ues for hundreds of ad airings in June, based on the minute of the hour when the ad was aired. Although almost all airings have %IAR values above 80%, the vast majorxity of the lowest-scoring airings occur at minutes 28, 29, 30, 58, and 59. Because these are also typical program boundaries, we have two explanations for this phenomena. First, many of the people tuning out at these minutes are doing so in search of new programs on other channels, not in response to a specific ad. Second, some of these low values may be attributable to DVR tuning, which also occur largely at program boundaries. But even after removing DVR events, the program-boundary effect is still visible, suggesting that the first explanation also holds true.

Figure 2 shows the susceptibility of %IAR to the size of the initial audience: as the initial audience increases the variance in %IAR decreases. Here we divided the airings into groups of equal number of airings by "low", "medium" and "high" initial audience and we show %IAR for evenings and mornings. Note that the variances decreases from "low" initial audience to "medium" and "high." On the other hand, the variance for a given audience size (e.g., "high") does not change significantly across different dayparts.

Figure 3 shows that different networks tend to have different characteristic %IAR measurements. Viewers seem to watch some networks more passively than others. Sports networks, for example, often have low %IAR measures on their ads, while children's networks tend to have high measures. This could be a reflection of the different audiences who watch these networks (and their characteristic viewing behavior), or it could be that the content itself lends itself to different styles of viewing[3].

Prior exposure to a given ad also seems to affect %IAR, often in a somewhat non-intuitive way. Figure 4 plots the %IAR of several hundred different ads aired in the month of August. Each time an ad is aired, the audience is divided up into first-time viewers, second-time viewers, etc., based on their previous exposures. We then calculate a %IAR for each of these sub-audiences. Figure 4 shows the average %IAR across all these ads, separated into sub-audiences in this way. The pattern is striking: the more often viewers have seen an ad over the last

[3]It could also be that the ads shown on some networks are simply more engaging than other ads, although the effect is so consistent by network genre that we consider this the least likely explanation.

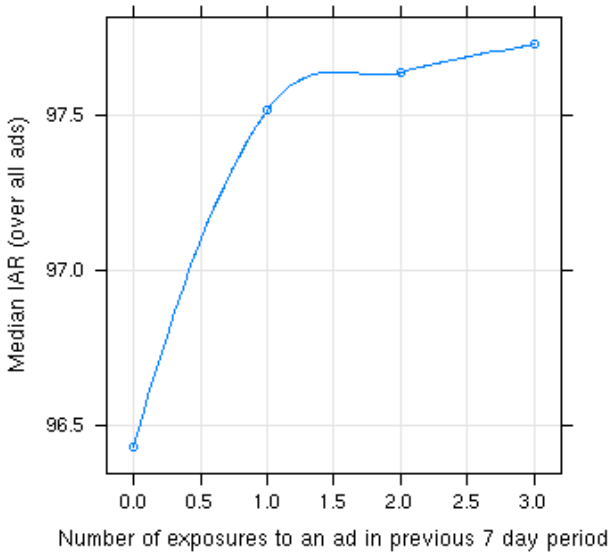

Figure 4: Impact of prior ad exposures on %IAR

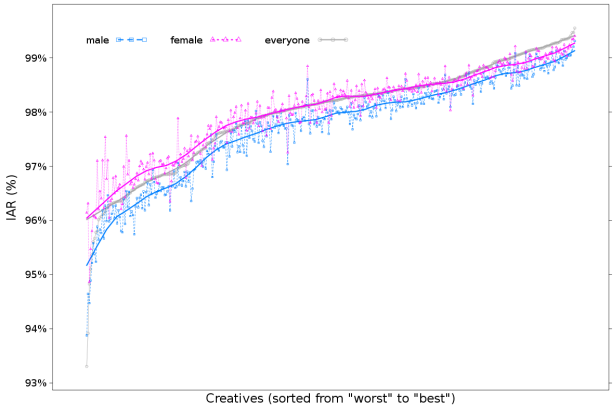

Figure 5: Gender differences in %IAR

month, the less likely they are to tune away[4].

We have recently begun exploring the ways demographics may also impact %IAR. For example, figure 5 plots the differing %IAR values calculated when looking only at households composed of a single female adult resident, a single male adult resident, and all households. Female households (pink triangles) consistently tune away less than male households (blue squares). The average %IAR across all households (gray circles) is generally somewhere in between these two, although for the ads with highest retention, both sets of single-adult households had lower-than-average audience retention[5]. We expect to find similar differences across viewers of differing ages and household incomes.

2 Measuring the creative effect

Using many of the results above, we have built statistical models for %IAR using daypart, network, pod position, ad duration, precise time, and day of week. The models attempts to predict the %IAR for a specific airing without knowing which creative will be run. We can then compare the actual %IAR we observe to this prediction. Ads that perform as expected are "normal," while ads that consistently deviate can be considered "good" or "bad" depending on which side of the prediction they fall on.

More precisely, we use the deviation from the model – the residuals – to rank creatives. We compute the fraction of the airings from a given creative that have residuals less than zero (underperforming airings), and then rank creatives using that fraction. A creative is deemed "bad" if at least 75% of its airings on a particular network are underperforming, and "good" if 75% of its airings are outperforming. We refer to these residuals as a "retention score," or sometimes, a "quality score."[6]

[4]To be clear, the causal direction here remains open to debate. It could be that prior viewership creates greater affinity for ads. But it could also be that more passive viewers are more likely to encounter ads multiple times. In other words, the cohort of single-exposure viewers may include many viewers who practice active ad avoidance, while the other cohorts contain fewer of these and so yield higher average retention.

[5]This may be due to the absence of children, by definition, in these households. As previously noted, we find that children's advertising appears to have especially high audience retention on average. This may be because children actually enjoy advertising more than adults, or, perhaps in some cases, because they cannot reach the remote control. We hope for the latter explanation.

[6]The term "ad quality" has a specific meaning in the con-

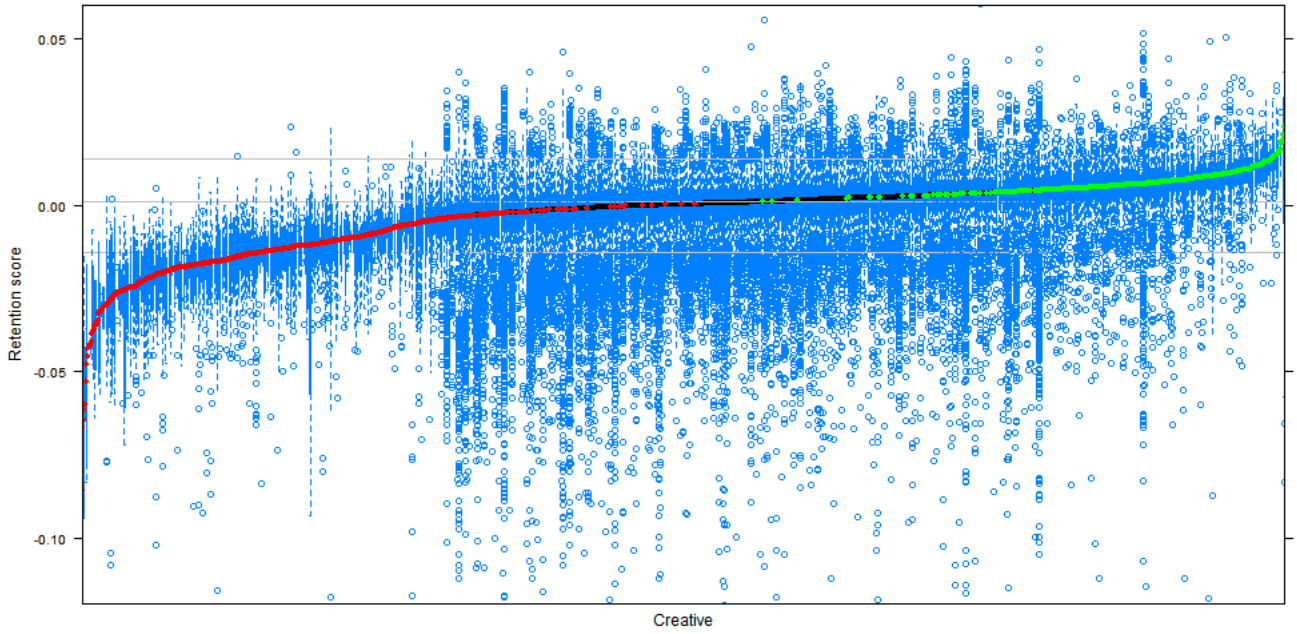

Figure 6: Distribution of residuals per creative.

Figure 6 shows the distribution of residuals per creative, with the creatives sorted by the median of their residuals. The creatives marked in red (on the left) appear to be underperforming (based on the 75% standard given above), while the creatives marked in green (on the right) are outperforming (based on the same standard).

To ensure that this rank is not an arbitrary artifact, we performed two cross-validation studies. We divided all the airings at random into two groups A and B. In figure 7, we plot points for every creative, showing the score across all airings in each group: the X axis is the fraction of residuals below zero for the airings from group A from the creative, and the Y axis is the same fraction corresponding to airings from group B. (We restricted the plot to creatives with at least 100 airings.) We see a strong correlation across the two random subsets.

In figure 8, we show a comparison of the ranking from month to month: the X axis here is fraction of residuals below zero for the airings from June, and Y axis is the same fraction corresponding to airings from July. The residuals for the two months are calculated from training data for that month. Again the chart shows a clear correlation, suggesting that our creative rank is stable.

3 Retention scores and human evaluation

In order understand further the meaning of the retention scores derived from the logistic regression described above, we conducted a simple survey of 78 Google employees. We asked each member of this admittedly unrepresentative sample to evaluate 20

text of Google's online advertising efforts that is generally associated with relevance: an ad with "high quality" is one that appears to be highly relevant to a given user in a given context. Although we sometimes use that term for historical reasons to describe our own work on television ads, we prefer the term "retention score," which described more precisely what is being measured and avoids the judgmental connotations of "quality."

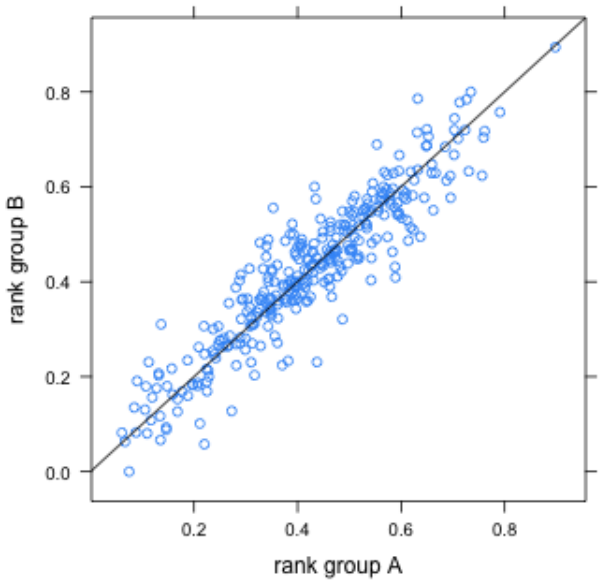

Figure 7: Comparing ranks per creative for models based on two random subsets of all ad airings.

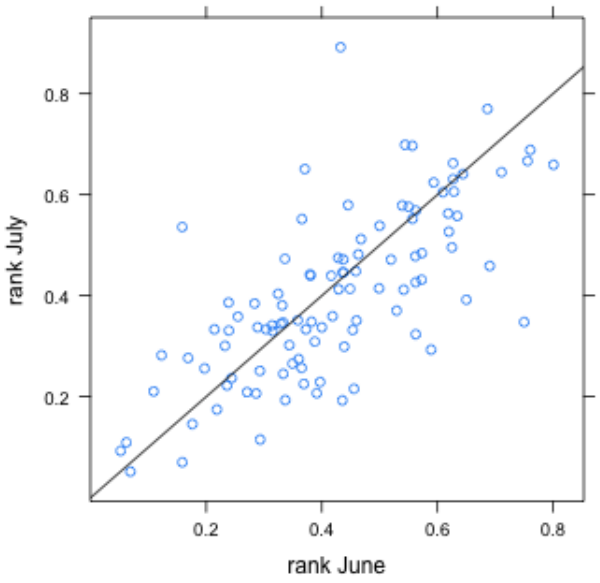

Figure 8: Comparing ranks per creative for ad airings from two consecutive months.

Human evaluation	Mean rank
At least "somewhat engaging"	14%
"Unremarkable"	38%
At least "somewhat annoying"	70%

Table 1: Correlating retention score rankings with human evaluations

television ads on a scale of 1 to 5, where 1 was "annoying" and 5 was "enjoyable." We chose these 20 test ads such that 10 of them had underperformed the model prediction in at least 75% of their airings (the so-called "bad ads"), and 10 of them had outperformed in at least 75% of their airings (the "good ads").

Table 3 summarizes the results. Ads that scored at least "somewhat engaging" (i.e, mean survey score greater than 3.5) averaged in the top 14th percentile of retention scores for all creatives. Ads that scored at the other end of the spectrum (mean less than 2.5) averaged in the 70th percentile. Ads with survey scores in between these two averaged in the 38th percentile.

Figure 9 gives another view of this data. Here the 20 ads are ranked according to their human evaluation, with the highest-scoring ads on top. The bars are colored according to which set of 10 they belonged to, with green ads coming from the group that outperformed the model and red ads coming from the group that underperformed. Although the correlation is far from perfect, we see fairly good separation of the "good" and "bad" ads, with the highest survey scores tending to go the ads with the best retention scores.

4 Predictive tests of retention scores

If retention scores based on the residuals shown in figure 6 are measuring some intrinsic property of the

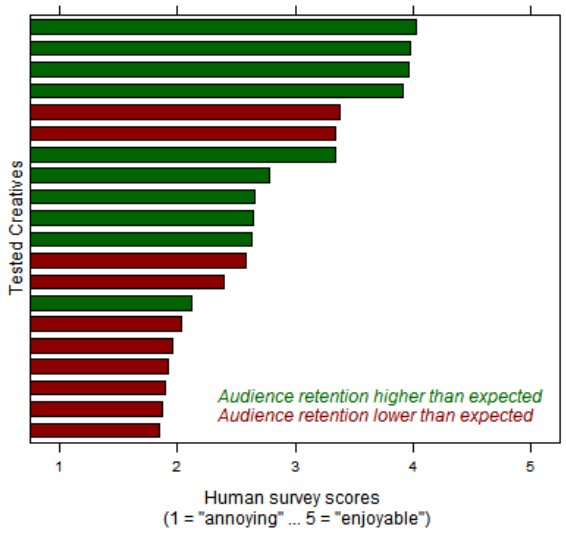

Figure 9: Correlating retention score rankings with human evaluations

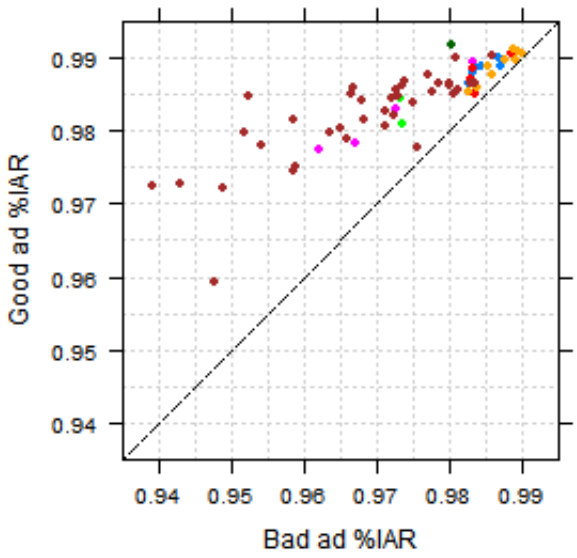

Figure 10: Predicting future relative %IAR based on retention scores

ads, then it should be possible to predict future audience behavior based on them. To test this, we selected pairs of "good" and "bad" ads and then ran these back-to-back on seven different TV networks[7] on several days between December 2008 and February 2009, for a total of 66 distinct airings. Because for each airing the non-creative factors (e.g., time of day, day of week, network, etc.) were held essentially constant[8], we would expect ads with positive retention scores to retain more audience than ads with negative retention scores.

Figure 10 shows the results of these 66 airings. The Y axis gives the %IAR for the "good" ad, while the X axis gives the %IAR for the "bad" ad. (The color of the points indicates different networks on which the ads were run.) Points above the diagonal line are those in which the "good" ad retained more audience. This was the case for all 66 airings, demonstrating that retention scores calculated from our model residuals are strong predictors of future ad performance.

These predictive tests represent the strongest evidence to date that our statistical models are able to isolate the impact of creatives on audience behavior, despite the significant noise introduced by non-creative factors.

5 Conclusions

Many factors influence the tuning behavior of TV audiences, making it difficult to understand the precise impact of a specific ad. However, by analyzing the tuning of millions of individuals across many thousands of ads, we can model these other factors and yield an estimate of the tuning attributable to a specific creative and confirm that creatives themselves do influence audience viewing behavior. This retention score — the deviation from the expected behavior — can be used to rank ads by their appeal, and perhaps

[7]The networks used were ABC Family, Bravo, Fine Living, Food Network, Home & Garden Television, The Learning Channel, and VH-1.

[8]We also alternated the order of the "good" and "bad" ads to neutralize any position bias.

relevance, to viewers, and could ultimately allow us to target advertising to a receptive audience much more precisely.

In the long run, we hope these methods will inspire and encourage more relevant advertising on television. Advertisers can use retention scores to evaluate how campaigns are resonating with customers. Networks and other programmers can use these same scores to inform ad placement and pricing. Most importantly, viewers can continue voting their ad preferences with ordinary remote controls — and using these techniques, we can finally count their votes and use the results to create a more rewarding viewing experience.