Lengthen your attribution window: Which digital ads have most long-term impact?

Received (in revised form): 28th February, 2023

Vivian Qin
Senior Data Scientist, Amazon Ads, USA

Vivian Qin received her doctoral degree in marketing from Fuqua School of Business, Duke University. She joined Amazon in April 2021. Currently, she is the team lead for data science within analytics and insights of Amazon Ads. She also advises the global research programme, and is the adjunct professor of marketing analytics at University of Maryland – College Park.

Amazon Ads, 410 10th Ave, 11th Floor, New York, NY 10001, USA
E-mail: qnyue@amazon.com

Abstract  Brands usually invest in a portfolio of digital ad products for brand consideration and conversion, and their performance is commonly evaluated with ad-attributed metrics. However, these metrics limit the measurement of advertising effectiveness within a short time window, typically of two weeks. Therefore, they could underestimate the total effect if some ad products’ efficacy lasts beyond the measurement period. In particular, this could understate the impact from ad products aimed at awareness and consideration. In addition, this bias could manifest in product categories where shoppers’ involvement is high because they are making deliberate purchase decisions. To solve these problems, the Vector Autoregressive Moving Average with Exogenous variables (VARMAX) model is applied, which allows flexibility in the length of the advertising measurement window, and thus can empirically quantify how long the effect of each ad lasts without a priori restrictions. For 15 US brands across three verticals (Hardlines, Softlines and Consumables) on Amazon, it was found that within the two-week attribution window, upper/middle-funnel ad products only materialise 30–50 per cent of the total effects, compared to lower-funnel at 60–90 per cent. Based on these results, it is recommended that advertisers and publishers lengthen the attribution window, and especially track their upper and middle-funnel ad products for at least a month to capture their longer-term effects.

KEYWORDS: digital ads, attribution window, performance metrics, long-term effects, e-commerce, ROAS

INTRODUCTION

‘Long Term Effects are essential for obtaining a fair evaluation of advertising effects.’ Raimund Wildner, Managing Director and VP, GFK Verein and Guido Modenbach, Managing Director of Seven One Media

Brands typically reference returns on advertising spend (ROAS) calculated with ad-attributed metrics to evaluate ad products’ effectiveness, and tend to do so within a short time period lasting 14 days or less.1 However, this short window may understate the importance of upper and middle-funnel
ads that improve awareness and consideration, as their effects may spill over to later stages of the consumer journey. Beyond delayed consumer behaviour, ad effects on sales could also last due to competitive response or company feedback effects. In fact, a 30-day attribution window is used to capture the long-term nature of business to business (B2B) marketing attribution. However, a one-month view is more of an exception than the norm. To further illustrate the longer-time effects of ads, previous literature suggests that display ads stimulate subsequent consumer visits through other advertising formats, such as visits increased from search. However, managers need to understand the long-term effects of all of their digital ad types, not just display, to optimise their allocation. Thus, this paper’s research question is: how long do the revenue effects of different digital ads last?

Digital advertising effectiveness is the topic of a rich academic literature but most has focused on the short-term effects of an ad campaign. Mass media ad effectiveness research has demonstrated that the carry-over and long-term effects of advertising differ significantly from its short-term effects. In her review of digital ad effectiveness, Liu-Thompkins therefore recommends as the first area for future research how the effect of an online ad exposure decays over time, noting preliminary research suggests that the rate of decay may differ by ad format.

This analysis uses data from 15 US brands that asked Amazon Ads to quantify ad carry-over effects. The carry-over effect of ad products is measured through a time-series model that allows lagged impact on the performance variables of revenue and detail page views (DPV), ie visits to the brand’s product pages on Amazon. This analysis generates three key findings. First, within the typical two-week attribution window, upper/middle-funnel ad products only materialise 30–50 per cent of the total effects, suggesting the insufficiency of the two-week measurement window. Secondly, contrasting with upper/middle-funnel products’ carry-over effect, most of lower-funnel ads’ (eg Sponsored Products) effects materialise within the attribution window, consistent with previous studies showing that lower-funnel products bring in immediate returns. Thirdly, when analysing across-vertical differences on customer consideration on DPV, the total effect within the 14-day window is the largest for Consumer Packaged Goods (CPG) (70 per cent), then Hardlines (45 per cent) and lastly Softlines (30 per cent). A likely explanation is that, regardless of ad products, shoppers in the low-involvement product categories (eg Consumables) will consider and implement purchases much faster than high-involvement categories (eg Hardlines and Softlines). This difference in the length of the consideration period could result in ad products in the former category showing peak performance much faster than the latter two.

Based on those findings, three recommendations are offered to advertisers. First, use Sponsored Products on search results pages for immediate returns as their impact is captured well within the two-week attribution window. Secondly, maintain a longer-term view for non-search ads (eg display and video ads) as only less than half of their effects realise within the existing attribution period, so their total impacts are underestimated. Finally, adjust the measurement window based on shopper involvement for a specific product category, as the same ad product could perform differently depending on the length of shoppers’ consideration period.

RESEARCH CONTEXT

Academic research has established that the long-term effect of advertising is about twice the short-term effect. Among the many reasons for such long-term effects, authors
Which digital ads have most long-term impact?

mention the longer purchase cycle of some products and services, the repeat purchases by ad-gained customers and the word-of-mouth and reviews by ad-gained customers, which can attract more customers. However, this empirical generalisation is based on TV and similar mass media advertising, and on operationalising the ‘short term’ as one year.\textsuperscript{12,13} However, current digital practice tracks an ad’s sales impact for a much shorter time, such as 14 days on Amazon, ignoring longer-term effects beyond this attribution window. This would not be a big issue if every ad type had a similar long-term/short-term return ratio. For example, advertisers could simply double each short-term effect to arrive at the long-term effect. However, the danger of misallocation arises if some ad types over-index on short-term impact, while others over-index on long-term impact.

This leaves us with the question of which digital ads should have a stronger long-term impact. Consistent with previous literature,\textsuperscript{14} it is proposed that this depends on the ad type’s goal in the purchase funnel/online decision journey.\textsuperscript{15} As Breuer and Brettle\textsuperscript{16} state:

Internet users have even more control over information processing than do readers of print articles: while they have on-demand access to a variety of information that is more diverse than any print item could ever be, they can control their level of advertising exposure (e.g. ignore the ad, click the ad, and interact with the ad by searching for ad related information).

Moreover, they assume that users of search engines are highly involved and that they remember paid search ads for a long time. Following this rationale, users of e-commerce sites such as Amazon should be even more involved in upper-funnel ads, because they browse the site to explore and buy products.\textsuperscript{17} Indeed, recent research showed longer lasting and stronger effects for content-integrated ads, ie ads that fit with the main reason the user browses the site.\textsuperscript{18} Note that a strong carryover effect does not mean that of the thousands of Internet users there is nobody who buys the advertised product immediately after seeing the ad.\textsuperscript{19} Still, many users will need time to carefully consider the information,\textsuperscript{20} especially for high involvement products.\textsuperscript{21}

**METHOD: PERSISTENCE MODELLING**

Marketing science literature has established the persistence modelling\textsuperscript{22} approach to examine carry-over effects in a flexible manner, ie without having to impose \textit{a priori} restrictions. The three steps involve univariate (one variable at a time), bivariate and multivariate tests.

The first step is unit root tests to establish that each variable is evolving (has a unit root) versus stationary (has a fixed average or baseline to which it returns). Stationary performance is the norm in mature markets, and implies that no marketing effects last forever (have persistence). Instead, they die out after some time, as estimated by the model. Likewise, stationary marketing variables go back to baseline after a change. Before including any variable in the model, it is necessary to ensure it is in fact stationary.

The second step is Granger Causality tests for each pair of variables (bivariate) to establish the direction of temporal causality: which variable moves before the other variable? While marketing is usually associated with causing sales, the opposite is also possible in a feedback effect of sales changes inducing marketers to act. Dual causality is rather common, and a key reason for estimating multiple equations models.

The third step is estimating the multiple equation model relating the variables to each other over time. In particular, the VARMAX (Vector Autoregressive Moving Average with X/Exogenous variables) model has three key benefits to assess long-term
advertising effects. First, the VARMAX model predicts the performance variables controlling for their own past (baseline) to obtain more precise estimates of the incremental impact of both current and past advertising. Secondly, it makes it possible to identify endogenous ad spend through the Granger causality test (ie those influenced by performance) and then allows that to influence performance in a two-way fashion. Lack of accounting for such feedback effect could bias ad measurement. Finally, the VARMAX model is more general than the Vector Autoregressive with Exogenous variables (VARX) that are nested within the VARMAX approach without moving average error terms. The twofold downsides of VARMAX are also acknowledged: 1) more parameters to estimate and thus larger sample is required; and 2) grid-search for the best model fitting hyperparameters can be computationally intensive.

The VARMAX model is generally specified as:

\[ y_t = \nu + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + Bx_t + \delta_t + M_1 \delta_{t-1} + M_2 \delta_{t-2} + M_q \delta_{t-q}, \]

where

- \( y_t \) is the multi-variate time series including both key performance indicators (KPIs) of interests and endogenous ad spend identified by the Granger Causality test, with \( p \) being the lag order (the Vector Autoregressive part of the VARMAX model)
- \( x_t \) is the exogenous ad spend identified by the Granger Causality test as not influenced by success metrics (the X part of the VARMAX model)
- \( \delta_t \) is the multi-variate error terms for \( y_t \), where \( q \) being the lag order (the Moving Average part of the VARMAX model)

The exact orders of \((p, q)\) are selected through grid-search based on the smallest forecasting errors (ie root-mean-square errors) of \( y_t \), in combination of Akaike Information Criterion (AIC) that penalises model complexity. Balancing between the above two methods enables the selection of a better-fitting model that does not over-parameterise and thus is less likely to overfit.

DATA

The 15 studied brands fall into one of the seven product categories across three verticals of Hardlines, Softlines and CPG: Electronics, Home & Kitchen, Tools & Home Improvement, Clothing, Shoes & Jewellery, Beauty & Personal Care, Health & Household and Grocery & Gourmet Food. Data for each brand ranges from one year to 1.5 years at the weekly level from January 2021 to June 2022. For each brand, tracking occurs of our two KPIs of DPV and revenue, and ad spend on Sponsored Products (SP), Sponsored Brands (SB), Sponsored Brands video (SBv), Sponsored Display (SD), Owned & Operated (O&O) Display, Demand-side-platform (DSP) Display and Streaming TV ads (STV). Various sources show that SP is the most preferred advertiser choice of search ads on Amazon, so it is defined as lower-funnel, ie closer to customer purchase decisions, relative to other types of ads.

MODEL TESTING

Stationarity. As the first step of persistence marketing, all KPIs are tested on stationarity by using with Augmented Dickey-Fuller (ADF) test. All tests reject the null hypothesis on the presence of a unit root with \( p-values <0.05 \). Thus all KPIs are classified as stationary without further differencing.

Granger Causality. The presence of feedback loop between independent and dependent variables is a key reason of using VARMAX models. To establish the necessity of current method, Granger Causality is tested on all possible combinations of time series. For each brand, KPIs cause ad investments for at least half of the combinations. In those cases, ad investments are included as endogenous in the
VARMAX model. In contrast, the exogenous ad spend (ie not influenced by KPIs) is the Exogenous/X part of the VARMAX model.

Order selection. For each brand, both grid search and the Akaike and Bayesian Information Criterion (AIB/BIC) are used to select the best fitting yet partimonious models.

Model fit. Three statistics are looked at to ensure acceptable model fit. First, Ljung-Box test is checked that detects the autocorrelations in error terms. All p-values are above 0.5 indicating the lack of residual autocorrelation. Secondly, heteroskedasticity is checked for by comparing the first third of the sample with the last third of the sample based on sum of squares. All have p-values above 0.2, suggesting no need to model heteroskedasticity. Thirdly, Jarque-Bera (JB) test is checked that is the equivalent of $R^2$ to measure goodness-of-fit in the sample. All have p-values above 0.2, indicating a good overall fit.

RESULTS
Insights are occurring by ad products and then category across all brands. Brand-level results are available in the Appendix.

Percentage of total effect calculation
After model fitting, ad products that have significant impact at selected lag orders are fitted with the impulse response function, which quantifies the impact of one shock (eg one unit of shock is defined as one standard deviation of the forecasting error) on KPIs of interests. As 15 brands are of different sizes in DPV and revenue, carry-over effect needs to be measured in a comparable way. So, for each unit shock of ad impressions by product types, the total effect on KPI is calculated up to seven weeks after the occurrence of shock. Then for each week, each week’s effect is divided by the total effect as the per cent of total effect realised over that week.

Ad product results
Though interaction effects between ads are tested, none is statistically significant. So for reporting purposes, each ad’s effect is shown standalone. Averaging across brand-level results, it is found that the upper-funnel ad product of DSP display shows about 40 per cent of its total revenue effect materialising in week five and beyond. For brevity, the summary for revenue is shown, since the DPV results are qualitatively similar (Figure 1).

Likewise, SD has a substantial part of its long-term revenue impact (~30 per cent) occurring after five weeks (Figure 2).

In contrast, SP does not show significant carry-over effects beyond three weeks for any brand. This is intuitive because SP is a lower-funnel ad product that aims at immediate returns on conversion.

Interestingly, STV ads’ effects are in between these two extremes, with carryover effects peaking in week two and three before declining. As a result, the percentage effect is more spread out over the weeks, with both week one and week five and beyond getting a large share of the overall effect. See Figure 3.

Category differences
Differences can also be observed in the degree of carry-over by category. For DPV, the largest percentage of carry-over beyond two weeks is in Softline (70 per cent) and the smallest for CPGs (30 per cent). Thus, the current attribution window of 14 days misses the majority of ad impact in Softlines, but a minority for CPG. For revenue, though the largest percentage of carryover beyond two weeks is still in Softline (75 per cent), the smallest becomes Hardlines (40 per cent). See Figures 4a to c.

It is inferred that regardless of the ad products, shoppers in the low-involvement product categories (eg Consumables) will consider and implement purchases much faster than the high-involvement categories...
Qin

ADSP Display’s Carry-Over Effect on Revenue

![ADSP Display's Carry-Over Effect on Revenue](image1)

Figure 1: Boxplot of percentage revenue effect of Amazon DSP (ADSP) Display by week. The boxplot shows the percentage of the effect’s distribution by week, where whisker on top is the maximum value and bottom the minimum value. The line in the middle of the box is the mean. In graphs, ‘week 5 and beyond’ has the highest average, i.e., across weeks, the largest effect on revenue materialises in week 5 and beyond. Note that these findings are based on five brands that invest in DSP Display.

SD’s Carry-Over Effect on Revenue

![SD’s Carry-Over Effect on Revenue](image2)

Figure 2: Boxplot of percentage revenue effect of SD by week. The boxplot shows the percentage of the effect’s distribution by week, where whisker on top is the maximum value and bottom the minimum value. The line in the middle of the box is the mean. In graphs, ‘week 5 and beyond’ has the highest average, i.e., across weeks, the largest effect on revenue materialises in week 5 and beyond. Note that these findings are based on eight brands that invest in DSP Display.

(eg Hardlines and Softlines), manifested through DPV. This difference in the length of shoppers’ consideration period could result in ad products in the former category showing peak performance much faster than the latter. The reversal of Hardlines and CPG of carryover effect in revenue, could be due to Subscribe and Save features used in Consumables. This feature enables automatic shipping and charging for customers signing up to receive products at regular cadence.
DISCUSSION

The analysis across 15 brands in different categories yields three main conclusions. First, upper/middle-funnel ad products only materialise 30–50 per cent of their total effects within a 14-day period, suggesting the insufficiency of the two-week attribution window. A longer-term view is recommended, such as one month to capture the majority of their impact on KPIs. Secondly, most of lower-funnel such as SPs’ success is captured within the standard attribution window. It is recommended that advertisers keep leveraging SP for immediate returns. Finally, carry-over effects beyond two weeks are vertical specific, with the smallest for CPG (30 per cent) and the largest for Softlines (70 per cent). It is recommended that advertisers in the high-involvement product categories (eg Softlines and Hardlines) hold a longer-term measurement view when gauging ads’ performances.

Would the findings generalise to other brands, categories and locales? The findings are indicative but they are based only on brands that asked for a long-term effects assessment. The results should be validated in a meta-analysis with more brands, especially brands that adopt STV, SBv and Video, to see if patterns discovered in this research continue to hold. Moreover, the analysis finds tentative evidence that the carry-over effect is vertical specific. Future research should explore more granular category differences. For example, within Hardlines, will advertising for big-ticket items like consumer electronics have an even longer carry-over effect than lower involvement products like tools for home improvement? Furthermore, for purchases that happen at regular cadence (eg Subscribe and Save), will shoppers be more deliberate and hence will advertising effects take longer to materialise?

As to limitations, the focus was on the duration and the size of the carryover effects in each period. This impulse response represents the net effect of all underlying factors, which include consumer behaviour, but also company feedback effects and competitive response. The current paper does not separate out these different explanations — it is simply reported that
Figure 4: Carry-over effect by product category
the net effects extend for several weeks, typically beyond the two-week attribution window.

**CONCLUSION**

Leveraging the sophisticated time series method of the VARMAX model, the relative effectiveness of ads in the longer term has been tested and empirically quantified better measurement windows depending on the length of the ad effect. With 15 brands included in the current study, it was consistently observed that within a two-week attribution window, upper/middle-funnel ad products only materialise 30–50 per cent of the total effects, compared to lower-funnel SP with the majority of the effect realised within the same window. Cross-vertical differences are also observed: CPG has the smallest carryover effect (30 per cent) beyond the 14-day window vs. Softlines (70 per cent). This suggests that the length of ad effects also depends on how consumers are involved in their purchases. It is recommended that advertisers and publishers lengthen the attribution window, and especially track their upper and middle-funnel ad products for at least a month to capture their longer-term effects.

**APPENDIX. BRAND RESULTS**

Table 1 shows the percentage of the total advertising effects occurring in week 1, 2, 3, 4, 5 and beyond for each ad product–brand combination. Note that some ad products, such as SPs, see all their effects dissipate after three weeks. In contrast, other products, such as DSP display, show a large part of their impact beyond the fourth week. These findings hold up for both KPIs, though the exact percentage differs for detail page views and revenues.

| Table 1: Percentage of Total Ad Effect Distributed Over Time on DPV (and Revenue)* |
|---------------------------------|---|---|---|---|---|---|
| **Week 1** | **Week 2** | **Week 3** | **Week 4** | **Week 5 and beyond** |
| **Brand A** | **Electronics** | **DSP Display** | 15% (17%) | 12% (5%) | 10% (6%) | 13% (14%) | 50% (58%) |
| | **SP** | 37% (48%) | 27% (23%) | 36% (29%) | 0% | 0% |
| **Brand B** | **Home and Kitchen** | **DSP Display** | 38% (35%) | 21% (14%) | 12% (12%) | 12% (19%) | 17% (20%) |
| | **SBv** | 70% (69%) | 30% (31%) | 53% (24%) | 0% | 0% |
| **Brand C** | **Tools & Home Improvement** | **DSP Display** | 0% (11%) | 47% (30%) | 33% (24%) | 3% (12%) | 47% (26%) |
| | **SB** | 2% (0%) | 19% (15%) | 28% (26%) | 17% (20%) |
| **Brand D** | **Electronics** | **SD** | 35% (35%) | 30% (32%) | 32% (31%) | 3% (2%) | 0% |
| | **SP** | 69% (62%) | 22% (26%) | 9% (12%) | 0% | 0% |
| **Brand E** | **Clothing, Shoes & Jewellery** | **DSP Display** | 0% (0%) | 17% (16%) | 7% (6%) | 29% (20%) | 47% (58%) |
| | **STV** | 14% (6%) | 39% (42%) | 322% (29%) | 11% (23%) | 14% (0%) |
| **Brand F** | **Home and Kitchen** | **DSP Display** | 53% (1%) | 12% (19%) | 3% (11%) | 7% (11%) | 24% (58%) |
| | **STV** | 11% (12%) | 25% (30%) | 34% (45%) | 14% (5%) | 16% (8%) |
| **Brand G** | **Beauty & Personal Care** | **O&O Display** | 81% (53%) | 4% (24%) | 7% (19%) | 7% (4%) | 0% |
| | **STV** | 80% (79%) | 14% (4%) | 2% (6%) | 2% (6%) | 2% (5%) |
| **Brand H** | **Tools & Home Improvement** | **SBv** | 70% (46%) | 30% (54%) | 0% | 0% | 0% |
| **Brand I** | **Health and Household** | **Video** | 22% (6%) | 31% (24%) | 32% (47%) | 15% (23%) | 0% |
| **Brand J** | **Home and Kitchen** | **DSP Display** | 13% (11%) | 24% (20%) | 24% (28%) | 10% (16%) | 29% (25%) |
| | **SD** | 28% (36%) | 17% (21%) | 1% (3%) | 2% (2%) | 52% (37%) |

(Continued)
Table 1: (continued)

<table>
<thead>
<tr>
<th>Brand</th>
<th>Product category</th>
<th>Ad product</th>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
<th>Week 5 and beyond</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand K</td>
<td>Home and Kitchen</td>
<td>SB</td>
<td>15% (11%)</td>
<td>26% (20%)</td>
<td>26% (28%)</td>
<td>11% (16%)</td>
<td>22% (24%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD</td>
<td>27% (35%)</td>
<td>18% (22%)</td>
<td>10% (10%)</td>
<td>5% (4%)</td>
<td>40% (29%)</td>
</tr>
<tr>
<td>Brand L</td>
<td>Grocery &amp; Gourmet Food</td>
<td>SD</td>
<td>8% (9%)</td>
<td>43% (28%)</td>
<td>19% (13%)</td>
<td>13% (18%)</td>
<td>17% (32%)</td>
</tr>
<tr>
<td>Brand M</td>
<td>Grocery &amp; Gourmet Food</td>
<td>DSP Display</td>
<td>40% (7%)</td>
<td>27% (8%)</td>
<td>12% (31%)</td>
<td>21% (5%)</td>
<td>0% (49%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SBv</td>
<td>6% (21%)</td>
<td>26% (25%)</td>
<td>23% (21%)</td>
<td>37% (2%)</td>
<td>7% (31%)</td>
</tr>
<tr>
<td>Brand N</td>
<td>Clothing, Shoes &amp; Jewellery</td>
<td>DSP Display</td>
<td>13% (14%)</td>
<td>19% (23%)</td>
<td>14% (19%)</td>
<td>20% (16%)</td>
<td>34% (27%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD</td>
<td>4% (4%)</td>
<td>16% (17%)</td>
<td>17% (21%)</td>
<td>16% (17%)</td>
<td>47% (27%)</td>
</tr>
<tr>
<td>Brand O</td>
<td>Home and Kitchen</td>
<td>STV</td>
<td>4% (10%)</td>
<td>8% (6%)</td>
<td>16% (13%)</td>
<td>23% (23%)</td>
<td>49% (48%)</td>
</tr>
</tbody>
</table>

* All shown effects are statistically significant at the p < 0.10 level, with the total effect adding up to 100% across weeks. For space limitation, ‘week 5 and beyond’ are grouped together though an individual week’s effect might still be significant.

References

13. de Haan, E., Wiesel, T. and Pauwels, K. (2016) ‘The Effectiveness of Different Forms of Online Advertising for Purchase Conversion in a Multiple-
19. Breuer and Brettel, see ref 10 above.

22. Dekimpe and Hansens, see ref 3 above.