OLIVER J. RUTZ, GARRETT P. SONNIER, and MICHAEL TRUSOV*

The authors propose a new approach to evaluate the perceptions and performance of a large set of paid search ads. This approach consists of two parts. First, primary data on hundreds of ads are collected through paired comparisons of their relative ability to generate awareness, interest, desire, action, and click performance. The authors use the Elo algorithm, a statistical model calibrated on paired comparisons, to score the full set of ads on relative perceptions and click performance. The estimated scores validate the theoretical link between perceptions and performance. Second, the authors predict the perceptions and performance of new ads relative to the existing set using textual content metrics. The predictive model allows for direct effects and interactions of the text metrics, resulting in a "large p, small n" problem. They address this problem with a novel Bayesian implementation of the VANISH model, a penalized regression approach that allows for differential treatment of main and interaction effects, in a system of equations. The authors demonstrate that this approach ably forecasts relative ad performance by leveraging perceptions inferred from content alone.

Keywords: electronic commerce, advertising, measurement and inference, Bayesian estimation, experiments

Online Supplement: http://dx.doi.org/10.1509/jmr.14.0186

A New Method to Aid Copy Testing of Paid Search Text Advertisements

In the last decade, paid search advertising has emerged as an important component of digital marketing. In 2015, total spending on Internet advertising amounted to $59.6 billion (PwC 2016). Of this overall spending amount, nearly half is related to search engine marketing. In paid search advertising, the search engine presents text ads to consumers in response to a keyword search. Although paid search text ads[1] share some similarities with more typical forms of advertising and promotion, there are some important differences in both form and function. Traditional advertising often aims to build awareness or create unique and favorable brand image associations. In contrast, the main goal of a text ad is to entice the consumer to respond immediately by clicking on the ad. In this sense, paid search ads share some commonalities with lower-funnel promotional tactics that aim to elicit a more immediate consumer response. However, an important difference between a text ad and, for example, a print advertisement or a point-of-sale promotional tool, is that the content of a text ad is restricted and, thus, very sparse. Text ads that appear on Google are formatted to consist of a headline, a display URL, and two lines of text (or a total of 95 characters).[2] This renders development of ad copy relatively inexpensive compared with most other forms of advertising.

Broadly speaking, firms might create ad copies for a paid search campaign using either human copywriters or machine-based approaches. An example of a machine-based approach is dynamic keyword insertion, in which the searched keyword is

[1] Hereinafter, we refer to paid search text ads as "text ads" or "ads."

*Oliver J. Rutz is Marion B. Ingersoll Professor and Associate Professor of Marketing, University of Washington (email: orutz@uw.edu). Garrett P. Sonnier is Associate Professor of Marketing and Zale Corporation Centennial Fellow, University of Texas at Austin (email: Garrett.Sonnier@mccombs.utexas.edu). Michael Trusov is Associate Professor of Marketing, University of Maryland (email: MTrusov@rhsmith.umd.edu). The authors would like to thank seminar participants at University of Southern California; University of Frankfurt; University of Colorado Boulder; INSEAD; Tuck School of Business; University of California, Riverside; University of California, Davis; University of Houston; and University of Utah for helpful comments. Coeditor: Robert Meyer; Associate Editor: Peter Danaher.

[2] Other search engines worldwide have similar ad formats. The leading search engine in China, Baidu, allows 20 Chinese characters for the headline and 100 Chinese characters for the body.

ISSN: 0022-2437 (print)
1547-7193 (electronic)

Journal of Marketing Research
Vol. LIV (December 2017), 885–900
DOI: 10.1509/jmr.14.0186

placed into a preexisting set of text to create an ad. In theory, this helps ensure the relevance of a text ad served in response to a keyword search while enabling advertisers to have small bodies of text (e.g., potentially a single ad) that appears differently to customers depending on their search terms. Whether created by humans or machines, it is relatively inexpensive to generate a large corpus of text advertisements. Copy testing these ads, however, is not as straightforward.

A common practice in advertising copy testing is to develop a set of copies (e.g., a few executions of a television or print ad) and experiment (e.g., using A/B split testing) to find the best-performing copies. Adapting this practice to copy test text ads is complicated by the scale of the problem. According to Google, the average click-through rate (CTR) on AdWords paid search ads is 2%. Under this average CTR, a copy test with only five ads that generates 20,000–30,000 impressions might take two to three months to generate enough data to yield reliable results (Rauschenbach 2012). In addition to the time required, if ads in the test group perform poorly in terms of generating clicks, the firm's quality score on the ad group may suffer, causing degradation in position, CTRs, and, ultimately, conversions for the whole campaign. Practitioners have noted this to be the case even in simple A/B tests (Rauschenbach 2012). Thus, although creating different text ads is relatively easy and inexpensive, testing a moderate to large set of text ads in a real business environment might be costly for the firm in terms of both time and money. It seems advantageous to consider methods to complement traditional copy testing by first reducing a large set of candidate text ads to a smaller set more suitable to in-market test procedures.

This article presents a novel approach to complement traditional copy testing methods (i.e., A/B or multivariate testing) in the domain of paid search advertising. From a broad vantage point, our approach consists of two parts. First, we demonstrate an experimental approach to scoring a large set of text ads on multiple dimensions of interest, including click performance and the perceived ability of the ad to generate awareness, interest, desire, and action. Second, we show how the textual characteristics of the ads can be used to predict the score of a new set of ads on these dimensions of interest. Our intended contribution is both substantive and methodological. From a substantive vantage point, our approach demonstrates how to identify a small set of higher-performing text ads from a large set of ads. This identification is facilitated by extant statistical methods to score a large set of objects on the basis of paired comparisons of objects in the set. Although our paired-comparison approach alone is easy to implement and can yield insights into ad copy performance, it is useful to also consider how to assess the perceptions and performance of new ads without rerunning the paired-comparison experiments. We demonstrate how the textual characteristics of the ads can be used to predict an out-of-sample ad's perceptions and performance scores. This prediction task raises a significant methodological challenge. The number of textual covariates is likely to be large, approaching or exceeding the number of ads (i.e., the "large p, small n" problem). Allowing for interaction effects aggravates the problem considerably. However, such effects are likely to be of interest in our setting. For example, the word "sale" and "now" may interact with each other if included in the same ad. To address this challenge, we propose a novel Bayesian approach to estimate a penalized system-of-equations model that addresses the "large p, small n" problem, allows for flexible treatment of main and interaction terms, and accounts for residual correlations.

We demonstrate our proposed approach using a set of 284 text ads.[3] First, we collect primary data on ad perceptions and performance through approximately 15,000 nonexhaustive paired-comparison experiments. Using a statistical ranking model to account for the nature of the paired-comparison data collection approach, we then rank-order the text ads in terms of the perceptions and performance scores. We find a significant relationship between an ad's perceived ability to generate interest, desire, and action and an ad's click performance, lending face validity to the proposed method. In addition, we repeat our approach using an auxiliary data set of text ads for which we observe actual click rates. Our derived click performance ranking correlates strongly with the ranking based on observed click rates, lending external validity to our proposed method.

To demonstrate the predictive validity of our approach, we specify a set of holdout ads from the set of 284 ads. Because we know the rank of each of the holdout ads, we can examine how well our approach forecasts the perceptions and performance of these ads relative to the existing set of ads. We calibrate the relationship between the textual content and perceptual attributes of the ads and use the model to forecast the holdout ads' perceptual attributes. We then predict performance on the basis of the perceptual attribute forecasts combined with ad content. This allows content to have a direct effect on performance as well as an indirect effect though the perceptions. Our approach ably predicts the rank of the holdout ads relative to the existing ads. We show that competing models, including those that do not account for interactions in textual content or that forecast performance directly as a function of content alone (i.e., bypassing the perceptual attributes), perform worse in terms of holdout performance.

The remainder of the article is structured as follows. We begin with a brief overview of the paid search advertising literature, including the treatment of the text advertisement. We then present a detailed discussion of our proposed method, including our approach to data collection. We follow with a description of the data and our empirical results. In the penultimate section, we discuss the implications of our framework and illustrate how new ads can be classified relative to the set of existing ads. We conclude by noting some limitations of our approach and discussing future research in the domain of text mining and marketing.

BACKGROUND

Designing an effective ad is perhaps one of the hottest topics among paid search practitioners and has been extensively discussed in numerous online forums. For example, a recent *Harvard Business Review* article discusses research conducted by eBay Research Labs (Blake, Nosko, and Taedelis 2015) concluding that paid search advertising has been largely ineffective in driving sales on eBay (Fisman 2013). Many in the practitioner community have argued that eBay's overreliance on dynamic keyword insertion has resulted in poor ad design and performance. Indeed, this strategy has resulted in the creation of some bizarre text advertisements suggesting that babies, perpetual motion machines, and even Bill Gates can be purchased on eBay for low prices (Kim 2013). The underlying theme of this and similar discussions among practitioners is that ad content in part determines ad performance (i.e., CTRs).

[3] Our results are reported for text ads from home lighting category. We replicated the results in two other product categories (for these results, see the Web Appendix).

Both industry experts and search engines offer numerous design recommendations. For example, Google AdWords has recommended "keep[ing] ad content simple"; focusing on unique features, such as price and promotional information; using a "strong call-to-action"; and including keywords in the ad text. Baidu has recommended embedding keywords to "help catch the user's eye" and to use plain language and clear descriptions, which "makes it easy for the user to understand what he/she will find on your landing page." While there are many trade publications and online sources that offer advice on how to design effective paid search ads, academic research on the effect of ad content is scarce.

Extant empirical research on paid search advertising has provided insights into keyword selection, the role of position and cost-per-click, and the general attribution problem (e.g., Chan, Wu, and Xie 2011; Ghose and Yang 2009; Rutz and Bucklin 2011). Dhar and Ghose (2010) provide an excellent overview of the state of paid search research and discuss future areas of research on the topic. In particular, this research has highlighted the effect of ad position in determining CTR as well as the importance of accounting for position endogeneity. Accounting for position endogeneity is especially challenging because finding strong, valid instruments is a difficult task (Rutz, Bucklin, and Sonnier 2012). As such, researchers often turn to field experiments or other methods (e.g., regression discontinuity designs) to evaluate paid search advertising (Agarwal, Hosanagar, and Smith 2011; Narayanan and Kalyanam 2015). Our goal is to understand whether and how subjective and objective measures of ad content can be used to predict click performance. Obtaining measures of the subjective perceptions of ad content requires us to collect primary data. Moreover, collecting primary data on click performance enables us to abstract away from the confounding effects of position and sharpen our focus on predicting performance as a function of content.

Computer science literature has studied the ability to predict the performance of a new ad. This stream of research on the role of text ads consists of three main areas: ad relevance to the search term or to the context (e.g., Broder et al. 2007; Gao and Gao 2013; Savenkov, Braslavski, and Lebedev 2011; Shaparenko, Cetin, and Iyer 2009), ad textual attributes (ad composition), and ad position (rank) on the result screen (Joachims et al. 2005). After controlling for position and relevance, studies have shown that large variation in ad performance remains (Richardson, Dominowska, and Ragno 2007). It seems reasonable that the textual content of the ad may account for some of this variation. Several papers in the computer science literature have studied the effects of the textual attributes of paid search ads. Kim et al. (2011) develop a factor graph model that links ad "attractiveness," measured by the attractiveness of individual words contained in the ad, to click-through performance. Leveraging historical performance records for a large set of ads, their model allows identification of words that are associated with better-performing ads. Dembczynski, Kotłowski, and Weiss (2008) propose a rule-induction algorithm approach to learn an ensemble of decision rules that help predict CTRs for new ads. Dave and Varma (2010) propose that semantically similar ads should have similar click-through performance. Their approach combines historical CTRs for a large set of existing ads with a similarity measure of new ads to generate predictions. Richardson, Dominowska, and Ragno (2007) use a logistic regression model to explore a broad range of ad attributes that can potentially be predictive of ad performance. Similar to our approach, they treat the ad copy as a "bag of words" and combine it with their own manually coded features (e.g., "aesthetically pleasing," "attention grabbing," "reputable brand"). Surprisingly, they find that neither the proposed ad features nor inclusion of interaction effects significantly improve the model's predictive accuracy.

From our perspective, a major difference between the aforementioned computer science studies and our work relates to the data. The majority of the articles we have discussed rely on data sets from search engines. In contrast, a typical advertiser does not have the luxury of observing contemporaneous or historical records of competitive search engine data. In essence, an advertiser has access only to its own data. Even if competitor data were available, it would not contain measures of the subjective perceptions of ad content (for either own or competitive ads). It is, however, relatively straightforward to capture the ads served in response to a search; the remaining task is to collect primary data on both consumers' subjective perceptions ad click rates. To this end, we next discuss our methodology in detail.

METHODOLOGY

We aim to achieve two distinct but related objectives with regard to understanding text ad performance (in terms of click behavior) and the role of content in determining the performance of a paid search campaign. First, we propose a novel market research approach to evaluate the relative performance of an *existing* set of text ads. Although in-market testing methods such as A/B and multivariate testing have the advantage of high external validity, they are not well-suited to test large sets of paid search ads, which are easy for copywriters or machine approaches to produce. Low CTRs mean that tests run with even a modest set of ads may need to stay in the field for months to accumulate enough data to yield reliable results. This may also result in negative effects on the firm's quality score if ads in the test group are performing poorly (Rauschenbach 2012). Our approach enables the firm to recover the rank order of a large set of ads in terms of perceptions and performance. This allows for the evaluation of a large number of text ads that may be served in response to a search. The firm may want to subject the resulting smaller set of high-performing ads to in-market testing or to roll out this smaller set of high-performing ads. In addition to collecting primary data on click performance through a paired-comparison exercise, it is also straightforward to measure consumers' subjective perceptions of the ad content. Following the growing literature that argues for the role of mindset metrics in understanding sales performance (e.g., Hanssens et al. 2014), we show that consumer perceptions explain ad performance and, more importantly, improve the prediction of the click performance for new out-of-sample ads, which is our second objective. Using only the new ad's textual content metrics and our model calibrated on existing ads, we are able to forecast the perceptions and click performance of the new ad. The advantage of this approach is that it obviates the need to collect primary data on the new ad. However, achieving this second objective raises some notable statistical challenges. Next, we discuss the motivation for a *relative* evaluation and then detail the methods for existing and new ads.

The Need for Primary Data Collection in a Relative (or Competitive) Setting

In an ideal setting, we could investigate the performance of text ads by collecting cross-firm CTRs for a set of ads served in response to a search. Several interesting considerations arise.

First, the ads served in response to a search are from a cross-section of firms. Second, since a set of ads is served, consumers are likely engaged in a *relative* evaluation (i.e., ad performance is affected by competitive context). Third, since the ads are typically served in a list, click performance is affected not only by the ad but also by the ad's position in the list. Thus, to evaluate the performance of text ads, the analyst likely would desire an approach that captures competitive click performance, provides insights into the relative perceptions and performance of a set of ads, and either controls for or abstracts away from the position effect. These considerations present significant challenges to the analyst. The major search engines do not provide competitive cross-firm data. Even if the analyst managed to obtain competitive click performance, measures of perceptions for own and competitive ads are not available. Moreover, even if the analyst overcame the two aforementioned limitations, existing research has demonstrated that ad position is endogenous, which complicates model-based efforts to control for position effects. Fortunately, the analyst can capture a set of competitive ads served in response to a search.[4] What is then required is obtaining relative measures of perceptual attributes (e.g., attracts attention, stimulates action) and performance (i.e., click) that are not confounded with position.

Evaluating Existing Paid Search Text Ads

The literature on classified advertising (i.e., ads that are similar to paid search ads in that they are also predominately text based) has suggested the importance of four perceptual attributes in understanding text ad performance: the ability of the ad (1) to generate *awareness*, (2) to stimulate *interest* in the product, (3) to stimulate *desire* for the product, and (4) to prompt consumer *to act* (Bruthiaux 1996, 2000; Vestergaard and Schroder 1985). Thus, our goal is to obtain relative measures on these perceptual attributes for a set of ads along with relative measures of performance (i.e., click). While it is possible to obtain aggregate measures of performance for a single firm using secondary data (i.e., average click behavior), neither competitive performance data nor data on the perceptual attributes are available through secondary data sets. Thus, we are motivated to collect primary data on perceptual attributes and performance. This task is complicated by the fact that we wish to investigate a large sample of ads on perceptions and performance. Standard marketing research methods to obtain perception and performance measures include ratings tasks, in which respondents might rate each ad on the dimensions (i.e., perceptions and click performance), or ranking tasks, in which respondents rank-order the set of ads on the dimensions. It is well known that ratings data suffer from low discriminant validity as well as scale usage biases. Ranking tasks can force discrimination across items but become cognitively more difficult as the number of objects to be ranked increases.

To address these issues, researchers in other settings have turned to the method of paired comparisons (MPC) and, more recently, maximum difference scaling (MaxDiff) to obtain measures on the dimensions of interest for each item. Given our desire to evaluate a large number of ads, designing MPC or MaxDiff experiments is problematic. Optimal experimental designs for MPC or MaxDiff require feature frequency balance, orthogonality, connectivity, and positional balance conditions (Sawtooth Software 2013). A rule of thumb for designing such experiments is to show one-and-a-half times as many paired comparisons as available items. Clearly, designing such experiments with many items (e.g., hundreds) is challenging. Both the sheer number of paired comparisons required per respondent and the financial costs of collecting such a large number of ratings are likely to be prohibitive.

Obtaining measures on a dimension or dimensions of interest for a large set of items is a common problem in the domain of sports. For example, there are 120 Division I college football teams, most of which never play each other during the 11–12 game college football season and certainly do not play each other according to any optimal experimental design. To determine postseason matchups, however, it is necessary to rank-order the teams on the basis of their performance in the season. Historically, subjective rankings of coaches and sportswriters determined the rank order of the top 20 teams. More recently, the Elo method (Elo 1978) has been used to rank-order all 120 teams. The Elo method is a rating system that can be used to rank items when exhaustive comparisons are not possible but paired comparisons are available and are not necessarily generated according to an optimal experimental design. The Elo method was developed by Arpad Elo for the purpose of rating chess players.[5] The basic premise of the method is that conditional on a certain rating of two items, a paired comparison of the items can be used to update their respective latent scores, s, which generate the rankings. The score update s_i^{new} for item i is based on the outcome of a paired comparison with item j as well as both items' precomparison scores, s_i and s_j. Thus, the updating step simultaneously takes into account the information contained in the outcome of the paired comparison as well as the information in the ratings of the items prior to the comparison. This is important because the comparisons are not necessarily generated (and may not be feasibly generated) according to an experimental design. For example, if a highly rated item is compared with a significantly lower-rated item, one would expect a higher probability that the highly rated item wins the comparison. Thus, not much new information is generated in this case, and both items' ratings will be updated only slightly. If, however, the significantly lower-rated item wins, considerable new information is discovered. Thus, both items' ratings should be updated to account for the fact that the two items are closer together than the prior ratings suggest.

The basic assumption of the original Elo method is that the underlying true rating of an item in a given comparison is a normally distributed random variable. In this sense, it is similar to random utility theory, wherein the analyst does not observe the consumer's utility but only the index of the utility-maximizing choice. Many implementations of the Elo method exist. We use an implementation called Rankr, developed by HP Labs (Luon, Aperjis, and Huberman 2010), which assumes a logistic distribution. The score update s_i^{new} for item i based on a paired comparison with item j, given their precomparison scores s_i and s_j, is given by

[4] For example, search result pages served by Google in response to a search can be automatically stored and parsed to extract the text ads served in response to a keyword search.

[5] Today, the Elo method is used to rate teams or individuals in other sports, such as basketball, Major League Baseball, and eSports.

$$s_i^{\text{new}} = s_i + \kappa\left[I(\text{choice} = i) - \frac{1}{1 + \exp(s_j - s_i)}\right], \tag{1}$$

where κ is the updating parameter and I is an indicator function. The Web Appendix contains more details on the Elo algorithm as well as a simulation study that demonstrates the efficacy of the approach.

Equation 1 can be calibrated on paired-comparison data collected on click performance for a set of ads (i.e., Given a pair of ads, on which ad would the respondent be more likely to click?). This gives the firm a sense of which ads in a large set are likely to have higher click rates. Likewise, Equation 1 can also be calibrated on paired-comparison data collected on the perceptual attributes for the same set of ads (i.e., Given the same pair of ads, which ad does the respondent rate more highly in terms of generating awareness or stimulating a desire to purchase?). After obtaining the Elo scores on click performance and perceptions, it seems prudent to investigate the relationship, suggested by theory, that an ad's perceptual attributes affect its performance. A model of ad performance as a function of perceptual attributes could be specified as follows:

$$s_i^{\text{Perf}} = \mu^{\text{Perf}} + \sum_{l=1}^{m} s_i^{\text{Att}_l}\theta_l + \omega_i, \tag{2}$$

where

s_i^{Perf} is the score of text ad i on performance,
$s_i^{\text{Att}_l}$ is the score of text ad i on the perceptual attribute dimension l (where l = 1, ..., m),
μ is an intercept,
$\theta = (\theta_1 \ldots \theta_m)'$ is a vector of parameters, and
ω_i is a normally distributed error term.

Such a model serves to assess the face validity of the Elo approach to ranking the ads on the basis of the paired comparisons. If we find no relationship between perceptual attributes and performance, this may suggest that the Elo method is unsatisfactory in terms of recovering the true relative performance and perceptions of the ads. Given the subjective nature of the perceptual attributes, it is especially worthwhile to estimate Equation 2 as one test of the validity of our approach. It is possible that measurement error can overwhelm any signal in the data. If this were the case, we would expect the estimates of θ to be attenuated.

Of greater practical consideration, should the firm want to assess the potential performance of new ads (which, in the domain of text ads, are easily and inexpensively generated), it would be required to collect new paired-comparison data with the original set of ads as well as the new ads. An alternative is to develop a model to predict the click performance of the new ads. This requires multiple steps. First, we need to measure the textual content of a set of ads through text mining. Then, we can build a predictive model of the perceptual attributes of an ad as a function of ad content. We can then use the estimates of θ from Equation 2 and the predicted perceptual attributes to generate a prediction of ad click performance. Because this is a complex approach, it is worth considering a simpler model that bypasses the perceptual attributes altogether and models click performance directly as a function of content alone. In either case, a problem arises in terms of the number of textual covariates relative to the number of observations.

Evaluating New Paid Search Text Ads: Linking Ad Content to Perceptual Attributes

Our proposed method is built on the notion that the perceptual attributes and, ultimately, the performance of the ad can be linked to the textual content of the ad. To capture the content of the text ads, we use multiple text-mining techniques. We begin by generating textual covariates using a variant of the bag-of-words method (Salton and McGill 1983) to capture the ad's content as well as the position of the word within the ad (header vs. body). The bag-of-words method is a simplifying representation used in natural language processing and information retrieval. It is an unordered representation of a text document determined by frequencies of words from a dictionary. Given the length (or the lack thereof) of text ads, words typically appear only once. Thus, rather than using word counts, we use word indicators. To account for higher-order structures in our data, we also use a latent Dirichlet allocation (LDA; Blei, Ng, and Jordan 2003) model to discover a mixture of latent topics present in the text ads. The LDA model is a generative statistical model that assumes that documents in the set (i.e., the set of 284 ads) are composed of multiple topics (i.e., collections of words constitute different topics). It has received widespread acceptance because of its power to extract meaningful topics using a parsimonious model. The LDA model operates by assuming that multiple latent topics exist in the text ads and that each topic is defined by a distribution of word probabilities. Although it is possible that there is some information overlap between the lower-dimensional topics and the bag-of-words covariates, we use both in our empirical analysis. In theory, the topic covariates contain the structured information in the ads, whereas the bag-of-words covariates represent the unstructured information.

While it is conceptually straightforward, modeling the perceptual attributes of an ad as a function of ad content captured by our text-mining approach will likely lead to a "large p, small n" problem (West 2003) because the number of textual covariates will be large relative to the number of text ads. Another complication arises as a result of the interest in interaction effects. Because text potentially becomes more meaningful as words co-occur, it is of interest to consider interactions. However, this implies that the "large p" grows even larger with interactions while the "small n" remains constant. Similar problems are often investigated in the statistics literature. In marketing, data sets are increasingly predictor-rich compared with the number of available observations (e.g., Naik et al. 2008). Thus, it is useful to consider methods to accommodate such data.

Regularization methods such as the LASSO (Tibshirani 1996) or Elastic Net (Zou and Hastie 2005) models offer means to address dimensionality problems in linear models. However, the consideration of interaction and main effects in the model gives rise to a second, related issue. In extant regularization models, interaction terms are treated simply as another covariate in the model. However, the researcher may prefer to add a main effect before an interaction. The researcher may also want to favor interaction terms that correspond to main effects that have already been added. We demonstrate how a novel regularization approach, the VANISH model (Radchenko and James 2010), allows for parsimonious treatment of the "large p, small n" problem in the presence of interaction terms.[6] We

[6] VANISH is the acronym for "variable selection using adaptive nonlinear interaction structures in high dimensions" (Radchenko and James 2010).

propose a Bayesian implementation of VANISH and derive the necessary informative prior required by this approach. A key advantage of a Bayesian approach to inference in the VANISH model is the ability to include the tuning parameters in the model hierarchy, allowing for simultaneous estimation of all parameters in the model (e.g., Park and Casella 2008). In addition, because we have multiple, possibly correlated perceptual attributes, we show how to generalize the VANISH model for a univariate dependent variable to a system model with correlated errors.

A Bayesian VANISH

In our analysis, the "large p, small n" problem, coupled with the issue of main and interaction effects, occurs in the step that links the textual covariates with the ad's perceptual attributes. For each text ad i (i = 1, ..., n), we have an Elo-based score[7] on each of the four perceptual attributes, $s_i^{Att_l}$; a set of textual covariates, X_i^{text}; and a set of topic covariates, Z_i^{topic}; from the LDA model. We model $s_i^{Att_l}$, text ad i's perceptual attribute score on dimension *l* (l = 1, ..., m), as a function of the topic covariates, the textual covariates, and the first-order interactions of the textual covariates[8] as follows:

$$(3)\qquad s_i^{Att_l} = \mu^{Att_l} + \sum_{t=1}^{nt} Z_{it}^{topic}\,\alpha_t^{Att_l} + \sum_{j=1}^{p} X_{ij}^{text}\,\beta_j^{Att_l} + \sum_{j>k} X_{ij}^{text} X_{ik}^{text}\,\beta_{jk}^{Att_l} + \varepsilon_i^{Att_l},$$

$$\varepsilon_i^{Att_l} \sim N\left(0,\ \sigma^2_{Att_l}\right),$$

where

n is the number of text ads,
m is the number of perceptual attributes,
nt is the number of topics in the LDA model,
p is the number of textual covariates, and
μ, α, β, and σ are parameters (or parameter vectors) to be estimated.

To implement the proposed model given in Equation 3, $(p^2 + p)/2$ beta parameters need to be estimated for each perceptual attribute dimension. Thus, estimating Equation 3 with standard methods of inference is simply not feasible for values of p that are large relative to the sample size. Furthermore, given our desire to consider differential treatment of main and interaction terms, we require an approach that accommodates such treatment. Next, we discuss the VANISH model and show how it allows for differential treatment of main and interaction effects. We then derive the informative prior needed to implement VANISH in a Bayesian framework and demonstrate how to extend the VANISH model to a system of equations with correlated errors.

Regularization methods such as VANISH impose penalties on the likelihood to account for the fact that the number of predictors is near or even larger than the number of observations. This circumvents the problems that would arise as a result of the lack of degrees of freedom. For a univariate linear model $y = X\beta + \varepsilon$, the general regularization approach is to find the value of $\hat{\beta}_p$ that minimizes $(y - X\beta)'(y - X\beta) - P$, the usual squared error criterion along with a penalty function, P, typically expressed in terms of β. In principle, standard regularization models can include interaction terms in the covariate matrix, X. However, extant models treat main effects and interactions in the same manner in terms of the penalty. We may prefer to ensure that main effects are included in the model before including their interaction term. In a similar vein, we may want to favor interaction terms corresponding to predictors that have already been added. The VANISH penalty automatically adjusts the degree of shrinkage on the interactions depending on whether the main effects are already present in the model. An added benefit of the VANISH penalty is the improved ease with which interaction terms can enter if the corresponding main effects have already been added. In general, the VANISH penalty for a given perceptual attribute dimension is

$$(4)\qquad P_{VANISH} = \lambda_1 \sum_{j=1}^{p}\left(\sum_{k=j+1}^{p}\left|\beta_{jk}\right|\right) + \lambda_2 \sum_{j=1}^{p}\left(\beta_j^2 + \sum_{k:k\neq j}^{p}\beta_{jk}^2\right)^{1/2},$$

where

λ₁ reflects the additional penalty on the interaction terms,
λ₂ reflects the weight of the penalty for each additional predictor included in the model,
p is the number of predictors, and
β are parameter vectors to be estimated.

Any Bayesian regularization approach specifies an informative prior on the otherwise-unidentified regression coefficients. In the case of the VANISH penalty, we derive the informative prior proceeds along similar conceptual lines as the LASSO.[9] The informative VANISH prior is given as

$$(5)\qquad \beta \propto \exp\left[-\frac{\lambda_1}{\sigma}\sum_{j=1}^{p}\left|\left(\sum_{k=j+1}^{p}\left|\beta_{jk}\right|\right)\right| - \frac{\lambda_2}{\sigma}\sum_{j=1}^{p}\left|\left(\beta_j^2 + \sum_{k:k\neq j}^{p}\beta_{jk}^2\right)^{1/2}\right|\right].$$

Unlike the LASSO, the form of the VANISH penalty does not allow for a normal posterior and requires a Metropolis–Hastings step to sample from the posterior (for details, see the Web Appendix). As noted by Park and Casella (2008), a benefit of the Bayesian approach is the ability to estimate the penalty parameters λ_1 and λ_2 directly, whereas in the classical framework, they are estimated using cross-validation. To complete the model, we specify gamma priors on the penalty parameters λ_1 and λ_2 and an inverse-gamma prior on the variance with diffuse hyper-priors (for more details on the sampler, see the Web Appendix).

As we have noted, the perceptual scores on awareness, interest, desire, and action might be correlated. However, given the informative prior for the univariate VANISH model specified in Equation 5, it is clear that a multivariate normal model with correlated errors is infeasible. We therefore use a copula approach to extend the VANISH model from a univariate dependent variable to a system of equations with correlated errors (Danaher and Smith 2011). For the ith text ad and lth attribute, we employ a Gaussian copula by first calculating (for each sweep of the sampler)

$$(6)\qquad x_{il}^{*} = \Phi\left[F_l\left(s_i^{Att_l}\right)\right],$$

[7] Hereinafter, we refer to the Elo-based score simply as the "score."

[8] The topic metrics already capture the whole text and thus are not interacted with the textual components or with themselves.

[9] For details on deriving the informative LASSO prior, see Park and Casella (2008).

where F is the normal cumulative distribution function with mean

$$\hat{m}_i^{Att_l} = \hat{\mu}^{Att_l} + \sum_{t=1}^{nt} Z_{it}^{topic}\,\hat{\alpha}_t^{Att_l} + \sum_{j=1}^{p} X_{ij}^{text}\,\hat{\beta}_j^{Att_l} + \sum_{j>k} X_{ij}^{text}\,X_{ik}^{text}\,\hat{\beta}_{jk}^{Att_l} \text{and variance}\left(\hat{\sigma}^2_{Att_l}\right).$$

Given the values x^*_{il} for all ads and attributes, we generate a new draw of the copula correlation matrix, Γ, using a random-walk Metropolis–Hastings algorithm. The copula parameters x^*_{il} and Γ allow us to model the perceptions conditional on the textual covariates in a multivariate system with correlated errors. The model is implemented in a fully Bayesian fashion and estimated by Markov chain Monte Carlo methods. More details on the sampler appear in the Web Appendix. After the model is estimated, it is straightforward to use the draws of the model parameters to forecast a new text ad's perceptual scores, conditional on the textual content of the ad.

Evaluating New Text Ads: Modeling Ad Performance

After obtaining a forecast of the new text ad's scores on the perceptual attributes, we can use the model in Equation 2 to predict the performance of the new text ad. Note that averaging over the draws of the Markov chain Monte Carlo sampler naturally accounts for the measurement error in the forecasts of the perceptual attributes. However, Equation 2 is myopic in the sense that it does not allow for a direct effect of ad content on performance. Our Bayesian VANISH approach can also be applied to allow for a direct effect of topic covariates, textual covariates, and first-order interactions of the textual covariates in the performance equation for the ith text ad:

$$(7)\quad s_i^{Perf} = \mu^{Perf} + \sum_{l=1}^{m} s_i^{Att_l}\theta_l + \sum_{t=1}^{nt} Z_{it}^{topic}\alpha_t^{Perf} + \sum_{j=1}^{p} X_{ij}^{text}\beta_j^{Perf} + \sum_{j>k} X_{ij}^{text}X_{ik}^{text}\beta_{jk}^{Perf} + \omega_i$$

$$\omega_i \sim N\left(0, \sigma^2_\omega\right),$$

where

Z_{it}^{topic} is the topic covariate t for text ad i,
X_{ij}^{text} is the textual covariate j for text ad i,
n is the number of text ads,
m is the number of perceptual attributes,
nt is the number of topics in the LDA model,
p is the number of textual covariates, and
μ, α, β, θ, and σ are parameters (or parameter vectors) to be estimated.

The model in Equation 7 is a more general model of ad performance. The model described in Equation 2 is derived as a special case of Equation 7, where $\alpha_t^{Perf} = \beta_j^{Perf} = \beta_{jk}^{Perf} = 0\ \forall t, j, k$. The restricted version of the model implies that ad content only affects performance through its impact on the perceptual attributes. Another restriction to consider is $\theta_1 = \theta_2 = \ldots = \theta_m = 0$, which implies that the perceptual attributes do not affect ad performance. By comparing the predictive validity of this restricted model with the full model, we can ascertain the value of collecting primary data on the perceptual attributes of the ad.

To summarize, our approach assembles a set of textual covariates that describe text ads using machine learning approaches. We augment these data with primary data that capture human processing of the text content on ad perceptions. Our Bayesian VANISH approach can generate forecasts of the perceptual attributes of a text ad given the ad's content alone. This is particularly useful in terms of forecasting the perceptions of a new text ad. More importantly, with a forecast of the perceptual attributes in hand, we can use Equation 7 in full or restricted form to predict a new ad's performance. Figure 1 provides a graphic description of our overall approach to establishing a link between the ad content, ad perceptions, and ad performance metrics relevant to practitioners (e.g., click behavior). To demonstrate the efficacy of our proposed approach, we now turn to our data.

DATA

Our data collection requires three steps. First, we need to collect a set of text ads to be evaluated. We then need to obtain measures of perceptions and performance for the ads. Finally, we need to obtain measures of ad content. We discuss each of these steps in turn.

Collecting the Ads for Evaluation

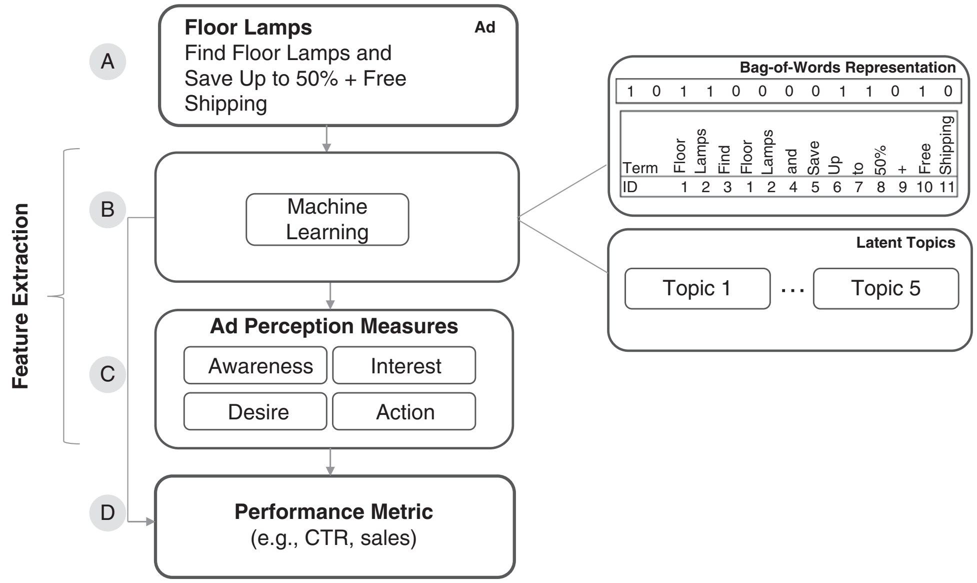
We obtained the ads used in our method from a mix of actual ads recovered in response to search and user-generated ads. To obtain the actual ads, we conducted a lab experiment in which approximately 300 participants were asked to perform a series of online Google searches and make purchase recommendations to a friend. Each participant was instructed to shop for a home lighting product (i.e., a halogen floor lamp).[10] Our participants used 215 distinct search phrases. However, most of these are simply different word orderings, the singular and plural of same terms, and misspellings. Overall, the search terms are quite homogeneous, and several of these seem to be autocompletes from the Google search engine. The phrase "halogen floor lamp(s)" accounts for 40% of the searches. Table 1 lists the top 20 terms which account for 67% of the searches. The remaining 195 terms account for the remaining 33% and are very consistent with terms shown here. The one exception of note is that there are some branded terms (i.e., IKEA, Home Depot, Target, Walmart, Amazon) that together account for 2% of the searches. As participants searched for products, the search result pages returned by Google were automatically recorded. The recorded pages were parsed and all paid search ads were extracted. After removing duplicate ads and ads from unrelated product categories, we obtained 140 ads for halogen floor lamps.

To illustrate the ease with which ad copy can be generated in the setting of paid search ads we use Amazon's Mechanical Turk service to (MTurk) create additional text ads. After signing up for our task on MTurk, respondents were taken to a page instructing them to create text ads for halogen floor lamps. We conducted data collection for this part of the study using a

[10] As noted in the introduction, we replicated our data collection procedure for two additional categories, GPS navigation systems and computer printers. The results, reported in the Web Appendix, are broadly consistent with those presented for the halogen floor lamp category.

Figure 1

OVERVIEW OF PROPOSED METHODOLOGY



proprietary web application automatically enforcing the format requirements of Google AdWords (i.e., 25, 35, and 35 characters for the header, line 1, and line 2, respectively). Our respondents generated 144 ads, resulting in a final count of 284 halogen floor lamp text ads. In our models of performance and perceptions, we include a dummy variable that captures whether the ad was an actual ad collected from Google versus generated by our respondents.

Collecting Perceptions and Performance Data

After the set of ads to be evaluated was determined, we collected data on ad perceptions and performance. As discussed previously, given the large numbers of items to be rated, traditional rate or rank methods are infeasible. Thus, we use a randomized system to serve a nonexhaustive set of paired ad comparisons whereby each respondent was shown 21 pairs.[11] We collected paired comparisons for perceptions and performance using MTurk. Figure 2 shows an example of a paired-comparison task measuring the relative ability of two ads to generate a desire to own the product. The wording of each paired-comparison task appears in Table 2. Drawing on our simulation study of the properties of the Elo method, we ensured that each text ad was shown in at least ten paired comparisons per dimension and that the pairs were picked at random (for details, see the Web Appendix). For the halogen floor lamps domain, we collected 15,242 paired comparisons, or, on average, 10.7 per text ad, per dimension. We then used the Elo method to recover the scores for the full set of ads in the domain (i.e., the ads are ranked on the basis of the Elo score) across the performance and perceptual attributes.

Measuring Textual Content

To capture the textual content of the ads, we employed multiple text-mining methods (for details, see the Web Appendix). First, we used the bag-of-words approach, arguably the most commonly used natural language processing method, to capture the textual elements of the ads. We used several common preprocessing techniques to create a set of predictor variables for our model. These include stemming as well as removing stop words, punctuation marks, and other special characters (Porter 1980). The best-performing set of 284 coded

Table 1

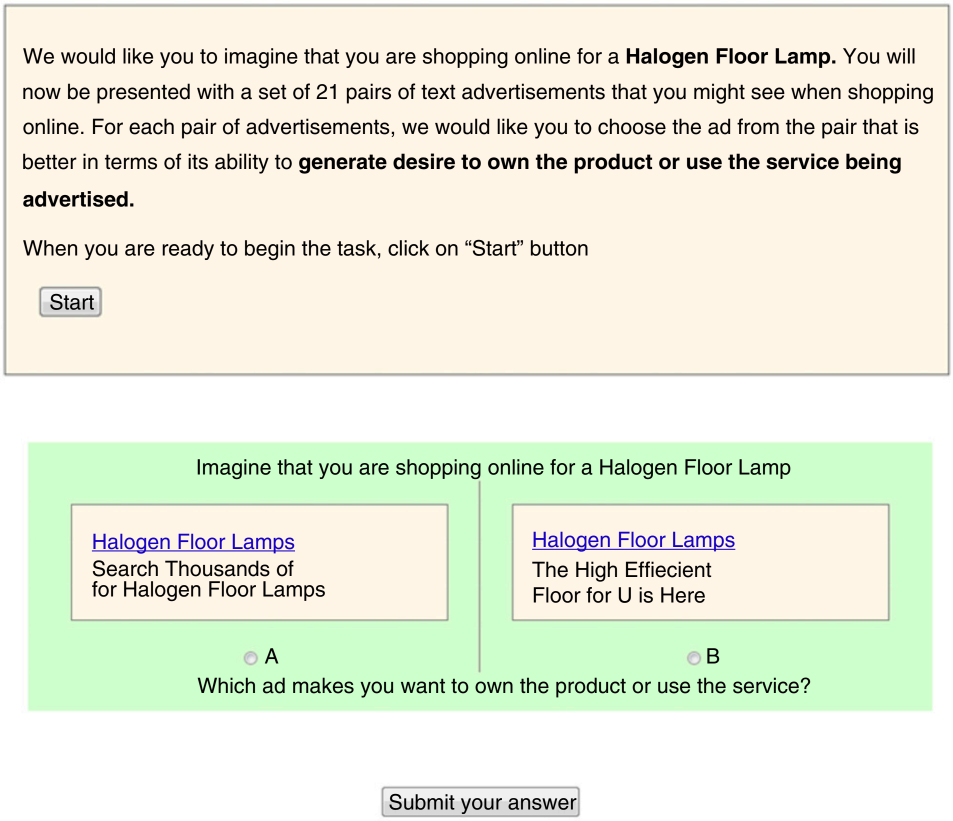
TOP 20 SEARCH PHRASES

Search Phrase	
halogen floor lamp	6ft 300 watt halogen floor lamp
halogen floor lamp 300 watt	halogen floor lamp cheap
300 watt halogen floor lamp	300 watt halogen lamp
halogen floor lamp 300w	6 ft 300 watt halogen floor lamp
halogen floor lamp 6 ft 300 watt	6 ft halogen floor lamp
halogen floor lamp 6ft	6ft Halogen floor lamp
halogen floor lamps	halogen floor lamp 300 watts
cheap halogen floor lamp	halogen floor lamp 6 feet 300 watt
halogen floor lamp 300 watt bulb	halogen floor lamp 6ft 300 watt
22halogen floor lamp 22	halogen floor lamp 6 ft

[11] A pretest showed that attention to the task decreases strongly after approximately 20 comparisons. Because respondents are paid per task, there was very limited interest in tasks that included only a small number of comparisons (e.g., 10). Balancing the need to attract respondents and the interest in quality data, we decided to use 20 pairs. We added a control pair after the 15th comparison. In this control pair, we had subjects compare a real ad with a control ad with grammatical and other obvious errors. Subjects who preferred the control ad were subsequently excluded from the data.

Figure 2

EXAMPLE OF PAIRED COMPARISON TASK



We would like you to imagine that you are shopping online for a **Halogen Floor Lamp.** You will now be presented with a set of 21 pairs of text advertisements that you might see when shopping online. For each pair of advertisements, we would like you to choose the ad from the pair that is better in terms of its ability to **generate desire to own the product or use the service being advertised.**

When you are ready to begin the task, click on “Start” button

Imagine that you are shopping online for a Halogen Floor Lamp

Halogen Floor Lamps
Search Thousands of
for Halogen Floor Lamps

Halogen Floor Lamps
The High Effiecient
Floor for U is Here

○ A

○ B

Which ad makes you want to own the product or use the service?

text ads for halogen floor lamps contained 308 unique words.[12] Because text ads consist of a header and body, we accommodated the different effects of word placement in the header versus body.[13] We augmented the 308 bag-of-words indicator variables by creating additional header/body indicator variables. Note that the same word can appear in both the header and the body. Our two-indicator header/body variable approach can capture this nonexclusivity feature. The text ads for halogen floor lamps averaged 11.5 words per ad. In our sample, the maximum (minimum) number of words per ad was 20 (5). Table 3 presents the top 20 words for the halogen floor lamp data. The top three words for halogen floor lamps are halogen (in 202 ads), floor (in 175 ads), and lamps (in 168 ads). We augmented the textual covariates with first-order interaction effects.[14] For halogen floor lamps, 1,087 interaction effects exist, resulting in a total of 1,683 textual predictors. With roughly 300 observations, a “large p, small n” problem arises when considering the relationship between perceptions and ad content represented by bag-of-words main- and first-order interaction textual covariates.

In addition to the bag-of-words approach, we also employed a topic model to discover the mixture of latent topics present in our sets of text ads. We used an LDA model to capture topics in our setting (Blei, Ng, and Jordan 2003). We fit the LDA model separately for our set of ads using five topic proportions that sum to unity. Table 4 reports the topic labels and representative words for the topics. In summary, we present the 284 text ads in our sample from a holistic vantage point using four predictors from the LDA topic model as well as from a more granular view using 1,683 indicator variables created by the bag-of-words approach, augmented with the location of the word (i.e., header vs. body) and first-order interactions. More information on our application of the LDA topic model is available in the Web Appendix.

RESULTS

We discuss our results in three parts. In the first part, we report on the Elo scores for click performance and perceptions for our 284 ads. We show that the perceptions are related to ad performance, which lends some face validity to our experimental approach. In the second part, we further demonstrate the validity of our approach. We report on the internal stability of the Elo rankings by computing multiple performance rank orderings of the same set of 284 ads derived from quasi

[12] For identification concerns, we included only words that appeared in at least two. For halogen floor lamps, 441 words appeared only in one ad. These are excluded from the analysis.

[13] We thank an anonymous reviewer for this helpful suggestion.

[14] Note that in our data, second-order (or higher-order) interactions are extremely rare because the text is constrained at 95 characters. Thus, we consider only first-order interactions.

Table 2

QUESTIONS FOR PAIRED COMPARISON TASK

Dimension	*Question*
Click	On which ad are you more likely to click?
Attention	Which ad is more attention getting?
Interest	Which ad makes you more interested in the product/service?
Desire	Which ad makes you want to own the product or use the service?
Action	Which ad motivates you to take some action to obtain the product/service?

bootstrapped sets of paired comparisons. We also report on evidence of the external validity of our experimental data collection procedure using a smaller set of auxiliary ads for which we observe actual click performance. In the third part, we report on the relationship between perceptions and content, which facilitates our effort to forecast the performance of a set of holdout ads. We compare our VANISH system model for predicting perceptions with competing system regularization models and show that our model provides the highest predictive validity. Using the forecasted perceptions from the models, we compute the predicted click performance of the holdout ads. Again, our VANISH system model provides the highest predictive validity. Finally, we show that a VANISH model of performance that includes both the perceptions as well as a direct effect of content outperforms a model that considers ad content alone (i.e., does not include the perceptions in the performance forecast equation).

Estimation of the Ad Performance and Perception Scores

An advantage of our experimental approach is that we are able to approximate click performance for an entire set of ads belonging to a domain (i.e., own and competitive ads) while abstracting away from the confounding effect of ad position. A second advantage of our experimental approach is the ability to measure and understand the perceived relative ability of a text ad to generate awareness, interest, desire, and action. We would expect to find positive correlations between the perceptual scores and the performance scores. Failure to find such a relationship may suggest that our paired-comparison approach to measuring perceptions and performance is not satisfactory. Figure 3 presents a matrix scatterplot of the Elo-based scores on click performance and perceptions for the halogen lamp data. Several insights emerge. First, Figure 3 shows that the performance and perceptions rankings are indeed positively correlated. Second, the relationship between performance and perceptions is approximately linear. Third, the correlations between the four perceptual attributes are not overwhelmingly high, indicating that multicollinearity should not be much of a problem in estimation of the performance equation (correlations range from .3 to .45). Our results

Table 3

TOP 20 WORDS

	Halogen Floor Lamps	
Rank	*Word*	*Count*
1	halogen	202
2	floor	175
3	lamps	168
4	lamp	95
5	!	87
6	your	60
7	of	54
8	light	47
9	free	46
10	for	43
11	and	42
12	on	41
13	the	41
14	shipping	40
15	&	39
16	to	37
17	prices	31
18	save	29
19	quality	28
20	bulbs	28

Table 4

LDA TOPICS AND REPRESENTATIVE WORDS

Topic	*Label*	*Representative Words*
1	Shipping	free, fast, shipping
2	Selection	lamps, quality, shop
3	Pricing	prices, low
4	Promotion	save, deals, bargain
5	Performance	light, easy, good

Figure 3

MATRIX SCATTERPLOTS OF ELO SCORES ON CLICK PERFORMANCE AND PERCEPTIONS

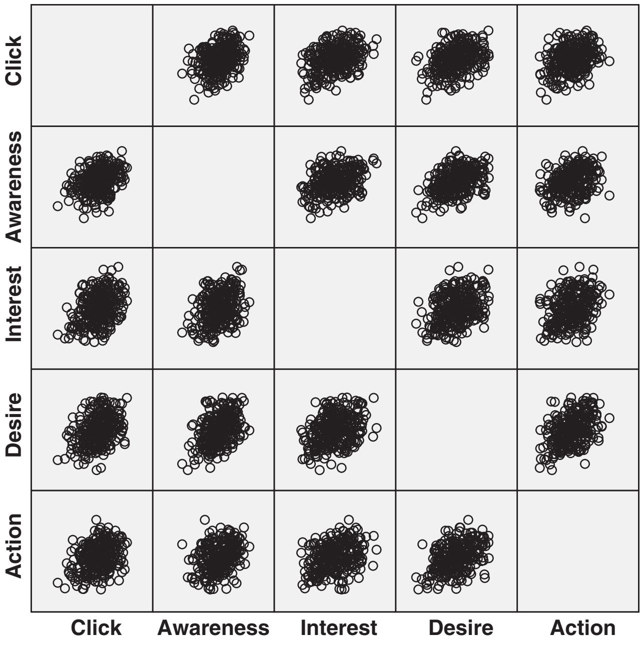


Table 5
TOP ADS BY PERFORMANCE AND PERCEPTION SCORES

Rank	*Header*	*Body*
Click		
1	Halogen Floor Lamp	100% Free Shipping! 100s of Halogen Floor Lamp Styles.
2	Buy Floor Lamps Online	Designer Brands at Discount Prices. Shop our Huge Selection and Save.
3	Crate & Barrel Floor Lamp	Shop Contemporary Floor Lamps for Your Home at Crate & Barrel Today!
4	Halogen Floor Lamps	Bargain Prices. Smart Deals. Save on Halogen Floor Lamps!
5	Halogen Floor Lamp	100s of Halogen Floor Lamps to Match Your Home. Free Ship Deals!
Awareness		
1	Floor Lamps on Sale Now	Floor Lamps for Less. Save 40% and Free Shipping - Order Online Today!
2	Floor Lamp Sale	Shop huge selection of Floor Lamps Free Shipping & Fast Delivery
3	Halogen Table Lamps	Exclusive styles for your home Ships Free Next Business Day!
4	Halogen Floor Lamps	New Halogen Lamps, Buy it on Sale 50%!
5	Best Halogen Floor Lamps	We Have Best Halogen Floor Lamps. Available in the Market, Try Them.
Interest		
1	Halogen Floor Lamps	Now Get your Own Floor Lamp for as Low as Never Before. Get Enlighted
2	Halogen Floor Lamps	Find Halogen Floor Lamps and Save Up to 50% + Free Shipping
3	Halogen Lamp at Low Price	Great Luminosity and Less Wattage. Saves Fixed and Running Cost
4	Buy Halogen Floor Lamps	Find 1000s of Floor Lamps to Match Your Style & Home. Free Ship Deals!
5	Floor Lamps At Lowe's	Find Stylish Floor Lamps Perfect To Light Your Rooms. Huge Selection!
Desire		
1	Save the Power	Less Power More Light. Make the Night as Day with Halogen
2	Halogen Floor Lamps	Find Halogen Floor Lamps and Save Up to 50% + Free Shipping
3	Halogen Floor Lamp	100% Free Shipping! 100s of Halogen Floor Lamp Styles.
4	HFL Brightness Glory	Eco Friendly & Power Conservation. Long & Brightful Life.
5	Halogen Floor Lamps	New Halogen Lamps, Buy it on Sale 50%!
Action		
1	Lamp Lighting	The Nation's Largest Specialty Lighting Retailer. Free Ship Deals!
2	Halogen Table Lamps	Exclusive styles for your home Ships Free Next Business Day!
3	Halogen Floor Lamp	100s of Halogen Floor Lamp Styles. Low Prices. Free Ship Deals!
4	Halogen Floor Lamps	Modern-Looking, Sleek and Versatile and Cheap
5	Halogen Floor Lamps	Very Bright. Save Power

suggest that our Elo-based approach for scoring the ads on perceptions and performance has good face validity.

As noted previously, our paired-comparison experiments enable the firm to identify the top ads in terms of click performance and perceptions. Table 5 presents the top five ads in terms of performance and perceptions for the halogen floor lamp data. A cursory examination of the header and body of the ads suggests that it is difficult to ascertain why certain ads do well in terms of performance and perceptions. Although our bag-of-words covariates are rich, they are also somewhat difficult, if not impossible, to interpret. However, the primary goal of the text-mining exercise is one of out-of-sample prediction, not causal inference (i.e., evaluation of ads, rather than a toolbox to create ads). Nonetheless, estimates of the relationship between ad performance, perceptions, topics, and text covariates as described in Equation 7 can shed some light on the drivers of ad performance. In particular, the LDA topics and the perceptual attributes lend themselves to interpretation.

Table 6 presents selected parameter estimates for the performance equation, including parameter estimates for the perceptions and the latent topics. We find positive and statistically significant relationships between the interest, desire, and action perceptual attributes and ad performance. Scores on awareness do not seem to explain variation in ad performance. These results are broadly consistent across the portable printer and GPS navigation data sets reported in the Web Appendix. Overall, the results suggest that ads that generate interest, stimulate desire, and call consumers to action perform better. This is consistent with the notion that paid search operates lower in the purchase funnel (i.e., the goal of search advertisements is generally not to create awareness). The LDA topics do not seem to add much explanatory power to our performance model. Only

Table 6
SELECTED PARAMETER ESTIMATES FOR PERFORMANCE EQUATION

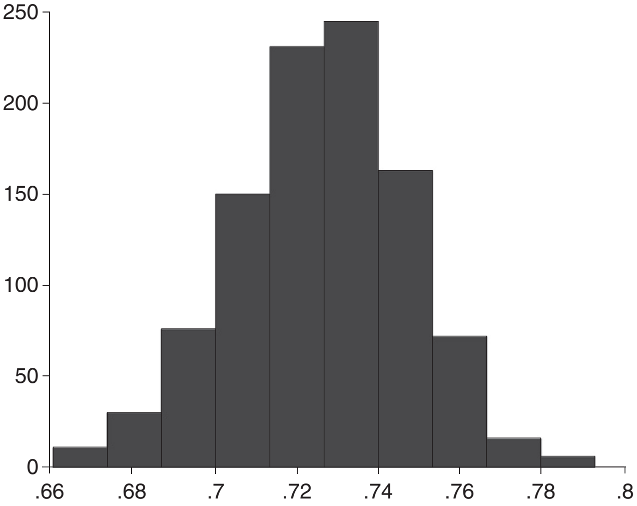
Parameter	*M (SE)*
Intercept	.15 (.26)
Topic 1: Shipping[a]	**−.17** (.05)
Topic 2: Selection	−.04 (.08)
Topic 3: Pricing	.11 (.06)
Topic 4: Promotion	−.04 (.08)
Awareness	−.15 (.08)
Interest	**.17** (.07)
Desire	**.28** (.07)
Action	**.19** (.07)
Ad dummy[b]	.11 (.07)

[a]Recall that the topics sum to 1; thus, we set the fifth topic (performance) to the baseline.

[b]Dummy variable captures actual versus user-generated ad, where user-generated ad is the baseline.

Notes: Boldface denotes that the 95% coverage interval does not include zero. In the interest of space, we do not report coefficient estimates for the 1,683 bag-of-word covariates.

Figure 4
HISTOGRAM OF RANK CORRELATIONS FOR 1,000 SAMPLES



one of the topic covariates is statistically significant. In terms of the performance difference between the actual and user-generated ads we find no effect for the lamp data. In contrast, the performance of the actual ads is higher than the user-generated ads for the replication data sets. In terms of effect sizes among the perceptions, changes in the desire score appear to have a larger impact on performance compared with changes in the interest and action scores. This result holds across all three data sets.

Validity Tests of the Experimental Approach

A concern regarding our paired-comparison data might be that measurement error overwhelms any signal in the data. However, estimates of the effect of the perceptions on ad performance across all three data sets suggest that this is not an issue. Pervasive measurement error would attenuate the estimates of the effect of perceptions on performance, which does not appear to be the case. However, additional concerns may arise over the internal stability of our experimental approach. To test the internal stability of the Elo ranking, we collected an additional sample of paired comparisons for our 284 text ads. The sample consists of 4,042 paired comparisons. For each paired comparison, respondents were asked to indicate on which ad they would be more likely to click. In a quasi-bootstrap approach, we created 1,000 randomly split samples. For each pair of samples, we computed the rankings using the Elo algorithm and computed the Spearman rank-correlation coefficient between the resulting two rankings. Figure 4 displays the histogram of correlations across the samples. The mean correlation is .72 with a standard deviation of .02. Thus, we conclude that our approach has high internal stability.

Concerns may also arise over the external validity of our experimental approach. To test external validity, we used an auxiliary, proprietary data set of 30 paid search ads tied to a single keyword from a mobile app category. For these 30 ads, we observed the actual click performance over a period of 60 days. This enabled us to rank the ads by observed click rate. Using this set of ads, we ran another pairwise comparison experiment. In this experiment, participants were only asked to indicate on which of two ads they would click (i.e., performance only). We generated 880 nonexhaustive pairwise comparisons to be rated by participants. From these pairwise comparisons, we calculated the ranking of the ads using Elo. We compared the Elo-based click ranking with the ranking based on observed CTRs. We find a Spearman rank correlation of .71 between Elo-based and observed rankings using raw CTRs and a Spearman rank correlation of .63 between Elo-based and observed rankings using position-adjusted CTRs (i.e., accounting for position).[15] We conclude that our approach has good external validity and our Elo-based click metric may be viewed as a good proxy for actual click behavior in the marketplace.

The Relationship Between Ad Content and Perceptual Attributes

We investigate the ad perception-content relationship on two related dimensions. First, should a model of perceptions based on content include main effects only or main effects and first-order interactions effects? Second, if first-order interactions effects are desirable, should they be treated like main effects, or should the modeling approach account for the relationship that exists between main and interaction effects? As previously noted, our goal is not to provide a tool box to create a successful ad.[16] Rather, we focus on a method that allows for the evaluation of new ads conditional on content in a setting where the content-related predictors are of high dimension relative to the number of ads. Consequently, we focus on the predictive performance of our proposed model versus three competing approaches. First, we use an elastic net model with main effects only to establish a baseline.[17] Second, we use

[15]To adjust for position, we regress the observed click rate on position and define the adjusted click rate as the observed click rate minus the expected click rate based on the regression parameter estimates.

[16]Such a tool box would need to not only indicate which words should be included but also account for grammar, sentence structure, and numerous other language peculiarities.

[17]The elastic net model is combination of Ridge and LASSO, alleviating the need to estimate the latter two models. We implement the elastic net using a Bayesian approach (Rutz, Trusov, and Bucklin 2011).

Table 7

MODEL COMPARISON: MSE OF HOLDOUT PREDICTIONS ON PERCEPTUAL DIMENSION

	Model			
	Elastic Net Without Interaction Effects	*Elastic Net with Interaction Effects*	*Intuitive LASSO*	*VANISH*
MSE: Perceptions	2.23	2.34	2.40	2.08

an elastic net model with main and interaction effects. Note that the elastic net model cannot discriminate between main and interaction effects and treats them in the same fashion. Third, we use an intuitive LASSO model that has one penalty structure for the main effects and a second penalty structure for the interaction effects. Although this model can penalize main and interaction effects differently, there is no built-in hierarchy between main and interaction effects that reflects the desire to prefer main effects over interaction effects. We compare these three alternative models with our proposed Bayesian VANISH model with main and interaction effects, treating main and interaction effects differentially in the penalty structure. In all cases, the perceptions are modeled as arising from a multivariate system conditional on the textual covariates with correlated errors accommodated through our copula approach.

To compare the models, we randomly selected 20% of our ads from the set of 284 to serve as a holdout sample. Our goal is to use the remaining 80% of the ads to calibrate the perception–content models discussed previously. Using the parameter estimates from the calibration sample, we then forecasted the holdout ads' perception scores and computed the mean squared error (MSE) of the forecast. Table 7 presents the results. Both the elastic net and the intuitive LASSO models perform worse than the elastic net model without interactions. From these results, one would erroneously conclude that when modeling the perception–content relationship, interaction effects do not need to be considered in the modeling approach. Our results indicate that this might be due to an incorrect treatment of interaction effects in the penalization structure. When penalizing main and interaction effects differently, a model that considers interactions outperforms a model without interactions. There is strong evidence that our proposed VANISH approach results in superior forecasting performance in terms of predicting perception scores for the holdout ads. The same patterns of results hold in our replication data sets. Table 8 presents the posterior mean estimates of the error correlations in the VANISH model estimated through our copula approach. The estimates show nonzero correlations across all dimensions of the system.

Forecasting Performance Based on Predicted Perceptual Attributes

Thus far, our proposed approach has enabled us to (1) provide a relative evaluation of text ads within a given (competitive) set on perceptions and performance, (2) measure the effect of perceptions on performance, and (3) ably forecast perceptions using ad content alone. Next, we integrate these pieces and show that our approach enables forecasting of performance by combining the insights from the performance–perception relationship with the forecasted perceptions. For each regularization model under consideration we generate the perceptual score forecasts and use these values along with the topic and bag-of-word covariates to predict ad performance as described by Equation 7. Note that we have the true rank for the holdout ads among the full set of ads from our paired-comparison task which we use to compute the MSE of the holdout prediction. The predictive validity of each approach is shown in Table 9. As with the perceptions data our system VANISH model provides the best out-of-sample performance forecast. This also holds across the replication data sets. To illustrate the value of measuring and utilizing perceptions we estimate a restricted version of Equation 7 for our best fitting VANISH model with the restriction that $\theta_l = 0$ for each of the perceptions. This reduces the performance model to a Bayesian VANISH model that bypasses the perceptions and models performance directly as a function of content. Table 10 shows that the model that uses the perceptions outperforms the restricted model that forecasts performance directly as a function of content. Again, this holds across the replication data sets.

Figure 5 presents a scatterplot of forecasted versus true ranks for the holdout ads. The Spearman rank correlation is .68. As the figure shows, the model is able to recover the three top ads (i.e., lower-left corner) very precisely. In addition, another ad is ranked very highly by the model, although this ad is ranked approximately 100th in the experiment. Thus, depending on the desired cutoff (e.g., all ads that the model suggests are top 30 ads), our model would have flagged three or four ads as high-performing ads that then can be tested in-market. Because two of these flagged ads are high-performing ads in our experiment, the model would have correctly identified these ads without the need to repeat the initial procedure of paired comparisons.

SUMMARY AND CONCLUSIONS

Paid search advertising is an increasingly large and important piece of firm advertising spending, accounting for nearly half of all Internet advertising spending. Much of the academic work on paid search has focused on understanding the relationship between ad position and click behavior, as well as the attribution of conversions to click behavior. These issues are undoubtedly important. However, it is intuitive to consider the role that the text ad itself may play in driving click-through performance. Paid search text ads share many

Table 8

ERROR CORRELATIONS FROM VANISH MODEL OF PERCEPTIONS

	Awareness	*Desire*	*Interest*	*Action*
Awareness	1.00	.35	.44	.42
Desire	.35	1.00	.32	.33
Interest	.44	.32	1.00	.45
Action	.42	.33	.45	1.00

Table 9

MODEL COMPARISON: MSE OF HOLDOUT PREDICTIONS ON PERFORMANCE

	Model			
	Elastic Net Without Interaction Effects	*Elastic Net with Interaction Effects*	*Intuitive LASSO*	*VANISH*
MSE: Click performance	1.31	1.45	1.55	.95

common characteristics with classified advertisements. In this domain, theory suggests that four important perceptual attributes drive the performance of an ad. Specifically, the ad's ability to generate awareness, stimulate interest and desire, and prompt consumers to act ultimately influence its performance. Indeed, when some paid search practitioners have questioned the efficacy of the paid search model, many have pointed out that the design of the ads used by the skeptics is ultimately to blame for poor performance. Presently, very little academic research has addressed this issue.

We present a novel modeling approach that enables managers to assess the relative performance of text ads viewed through the lens of these perceptual attributes. While managers often use A/B or multivariate testing to evaluate different paid search ad copies, such approaches are not well-suited to evaluate large numbers of ads. Practitioners have noted that even with a relatively small number of ads (e.g., five ads), such tests might take two to three months to generate enough data to yield reliable results and may adversely affect the firm's quality score on the ad group, causing a degradation in position, CTRs, and, ultimately, conversions. Our approach complements A/B and multivariate testing by effectively prescreening a large number of ads, and it yields a ranking of the set of ads on perceptions and performance based on paired comparisons. Importantly, our approach does not require that the comparisons be generated according to any experimental design. In some cases, researchers can leverage the MPC where the pairs are served to respondents according to an optimal experimental design. However, such methods are not feasible when the number of objects to be evaluated is large. To solve this problem, we leverage methods used in sports to rank-order hundreds of players and teams when only nonexhaustive paired comparisons are available and the comparisons are not generated by an experimental design. We show demonstrate this method ably rank-orders the ads in terms of perceptions and performance. Furthermore, we show that the scores generated by this approach are internally stable, closely approximate observed click behavior for paid search ads and are consistent with theory that suggests a positive relationship between perceptions and performance.

Our approach to score a large set of existing ads on perceptions and performance is straightforward for managers to implement. However, if the manager wants to evaluate a new ad in relation to the existing set, (s)he would be required to rerun the experiment. To overcome this limitation, we show how to forecast the performance of a new ad on the basis of the ad's observed textual content alone. We capture the textual content of a set of ads using techniques from the natural language processing and information retrieval fields. To compute the performance forecasts, we first forecast perceptions as a function of textual content. The forecasted perceptions and content metrics can then be used to forecast an ad's performance using the results from the calibration of the performance–perception relationship. Alternatively, we also can bypass perceptions and forecast performance directly and solely as a function of content.

Although our approach to forecasting ad performance is, in principle, straightforward, a significant challenge arises because the number of textual covariates is quite large relative to the number of ads. Furthermore, the co-occurrence of textual covariates is likely of interest. Investigation of this issue aggravates the problem of accounting for a large number of predictors given a relative small set of observations. Some common penalized regression approaches are available to address this so-called "large p, small n" problem but do not distinguish between main and interaction terms in the penalty function. We utilize a novel penalized regression approach that differentially penalizes main and interaction terms. We extend the model to a multivariate setting and demonstrate a Bayesian approach to estimation. We apply our model to the problem of predicting the performance of a set of holdout ads and show that our approach outperforms existing penalized regression

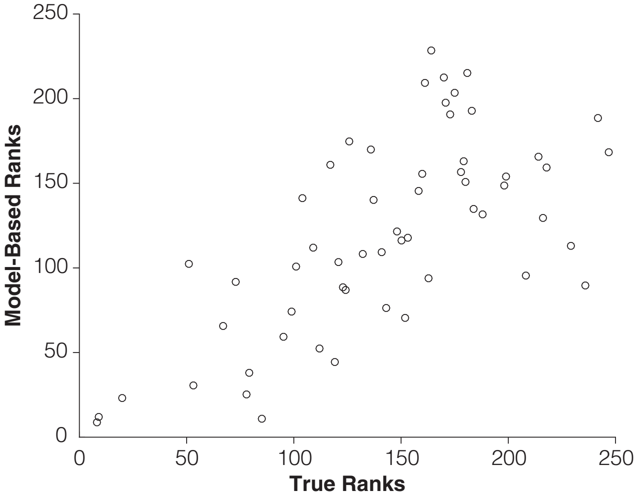
Table 10

VANISH MODEL COMPARISON: MSE OF HOLDOUT PREDICTIONS ON PERFORMANCE

	Model	
	VANISH with Perceptual Attributes	*VANISH Without Perceptual Attributes*
MSE: Click performance	.95	1.16

Figure 5

PREDICTED VERSUS ACTUAL TEXT AD PERFORMANCE



approaches in terms of predicting the perceptual attributes of a holdout set of ads on the basis of ad content. We also demonstrate that our proposed approach outperforms a model that disregards the role of perceptions in forecasting text ad performance (i.e., a model of perceptions solely as a function of textual content).

Overall, our approach provides managers with a tool to assess relative perceptions and performance of a large set of text ads and, importantly, to forecast perceptions and performance of new ads on the basis of ad content alone. It is, however, not without limitations. Similar to more traditional MPC approaches (or even rate or rank tasks), our approach identifies relative perceptions and performance but cannot speak to absolute perceptions or performance. In terms of quantifying the content of the ad, words that occur only once are removed from the analysis. Although we focus more on generic search, this may be a limitation in a branded search context if such words occur once because of their uniqueness to a specific retailer. An implicit assumption is that the new ads can be described by the same dictionary of words and topics as the calibration ads. Furthermore, the same word or words may take on different meanings in another context. While our interactions and LDA topics capture context to some extent, further refinements in text analytics may improve on our ability to quantitatively represent text. Similar to machine learning techniques, our VANISH model is well-suited to prediction in an underdetermined system. However, drawing substantive insights from parameter estimates is a challenge. A broader limitation of our approach is due to the dynamic nature of paid search advertising. For example, the attractiveness and performance of an ad may change dynamically as a result of various factors. Accounting for such dynamics in an experimental setting such as ours is a considerable challenge. Future research may consider how to overcome some or all of these limitations.

REFERENCES

Agarwal, Ashish, Kartik Hosanagar, and Michael D. Smith (2011), "Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets," *Journal of Marketing Research*, 48 (6), 1057–73.

Blake, Thomas, Chris Nosko, and Steven Taedelis (2015), "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment," *Econometrica*, 83 (1), 155–74.

Blei, David M., Andrew Ng, and Michael I. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Broder, Andrei, Marcus Fontoura, Vanja Josifovski, and Lance Riedel (2007), "A Semantic Approach to Contextual Advertising," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: Association for Computing Machinery, 559–66.

Bruthiaux, Paul (1996), *The Discourse of Classified Advertising: Exploring the Nature of Linguistic Simplicity*. Oxford, UK: Oxford University Press.

Bruthiaux, Paul (2000), "In a Nutshell: Persuasion in the Spatially Constrained Language of Advertising," *Language & Communication*, 20 (4), 297–310.

Chan, Tat Y., Chunhua Wu, and Ying Xie (2011), "Measuring the Lifetime Value of Customers Acquired from Google Search Advertising," *Marketing Science*, 30 (5), 837–50.

Danaher, Peter J., and Michael S. Smith (2011), "Modeling Multivariate Distributions Using Copulas: Applications in Marketing," *Marketing Science*, 30 (1), 4–21.

Dave, Kushal, and Vasudeva Varma (2010), "Learning the Click-Through Rate for Rare/New Ads from Similar Ads," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: Association for Computing Machinery, 897–98.

Dembczynski, Krzystof, Wojciech Kotłowski, and Dawid Weiss (2008), "Predicting Ads' Click-Through Rate with Decision Rules," *TROA: Workshop on Targeting and Ranking for Online Advertising*, https://pdfs.semanticscholar.org/0019/f17bb98c72a9ce24a67fc903f969b8e261c6.pdf.

Dhar, Vasant, and Anindya Ghose (2010), "Sponsored Search and Market Efficiency," *Information Systems Research*, 21 (4), 760–72.

Elo, Arpad (1978), *The Rating of Chessplayers, Past and Present*. New York: Arco Publishing.

Fisman, Ray (2013), "Did eBay Just Prove That Paid Search Ads Don't Work?" *Harvard Business Review* (March 11), http://blogs.hbr.org/cs/2013/03/did_ebay_just_prove_that_paid.html.

Gao, Zhe, and Qigang Gao (2013), "Ad-Centric Model Discovery for Predicting Ads' Click-Through Rate," *Procedia Computer Science*, 19, 155–62.

Ghose, Anindya, and Sha Yang (2009), "An Empirical Analysis of Search Engine Advertising: Sponsored Search," *Management Science*, 55 (10), 1605–22.

Hanssens, Dominique M., Koen H. Pauwels, Shuba Srinivasan, Marc Vanhuele, and Gokhan Yildirim (2014), "Consumer Attitude Metrics for Guiding Marketing Mix Decisions," *Marketing Science*, 33 (4), 534–50.

Joachims, Thorsten, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay (2005), "Accurately Interpreting Clickthrough Data as Implicit Feedback," in *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: Association for Computing Machinery, 154–61.

Kim, Larry (2013), "Dear eBay, Your Ads Don't Work Because They Suck," *WordStream Blog AdWords Tips* (May 12), http://www.wordstream.com/blog/ws/2013/03/13/dear-ebay-its-not-adwords-its-you.

Kim, Sungchul, Tao Qin, Hwanjo Yu, and Tie-Yan Liu (2011), "An Advertiser-Centric Approach to Understand User Click Behavior in Sponsored Search," in *CIKM '11 Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. New York: Association for Computing Machinery: 2121–24.

Luon, Yarun, Christina Aperjis, and Bernado A. Huberman (2010), "Rankr: A Mobile System for Crowdsourcing Opinions," in *Mobile Computing, Applications, and Services. MobiCASE 2011: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, Vol 95, J.Y. Zhang, J. Wilkiewicz, and A. Nahapetian, eds. Berlin: Springer, 21–30.

Naik, Prasad, Michel Wedel, Lynd Bacon, Anand Bodapati, Eric Bradlow, Wagner Kamakura, et al. (2008), "Challenges and Opportunities in High-Dimensional Choice Data Analyses," *Marketing Letters*, 19 (3), 201–13.

Narayanan, Sridhar, and Kirthi Kalyanam (2015), "Position Effects in Search Advertising and Their Moderators: A Regression Discontinuity Approach," *Marketing Science*, 34 (3), 388–407.

Park, Trevor, and George Casella (2008), "The Bayesian LASSO," *Journal of the American Statistical Association*, 103, 681–86.

Porter, Martin F. (1980), "An Algorithm for Suffix Stripping," *Program*, 14 (3), 130–37.

PwC (2016), "IAB Internet Advertising Revenue Report," research report, https://www.iab.com/wp-content/uploads/2016/04/IAB-Internet-Advertising-Revenue-Report-FY-2015.pdf.

Radchenko, Peter, and Gareth M. James (2010), "Variable Selection Using Adaptive Non-Linear Interaction Structures in High

Dimensions," *Journal of the American Statistical Association*, 105 (492), 1541–53.

Rauschenbach, Brian (2012), "Paid Search Ad Copy Auditing," *Whiteboard Friday* (February 9), https://moz.com/blog/paid-search-ad-copy-auditing-whiteboard-friday

Richardson, Matthew, Ewa Dominowska, and Robert Ragno (2007), "Predicting Clicks: Estimating the Click-Through Rate for New Ads," in *Proceedings of the Sixteenth International World Wide Web Conference*. New York: Association for Computing Machinery, 521–29.

Rutz, Oliver J., and Randolph E. Bucklin (2011), "From Generic to Branded: A Model of Spillover in Paid Search Advertising," *Journal of Marketing Research*, 48 (1), 87–102.

Rutz, Oliver J., Randolph E. Bucklin, and Garrett P. Sonnier (2012), "A Latent Instrumental Variables Approach to Modeling Keyword Conversion in Paid Search Advertising," *Journal of Marketing Research*, 49 (3), 306–19.

Rutz, Oliver J., Michael Trusov, and Randolph E. Bucklin (2011), "Modeling Indirect Effects of Paid Search Advertising: Which Keywords Lead to More Future Visits?" *Marketing Science*, 30 (4), 646–65.

Salton, Gerard, and Michael J. McGill (1983), *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Savenkov, Denis, Pavel Braslavski, and Mikhail Lebedev (2011), "Search Snippet Evaluation at Yandex: Lessons Learned and Future Directions," in *Multilingual and Multimodal Information Access Evaluation. CLEF 2011: Lecture Notes in Computer Science*, Vol 6941, P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, and M. de Rijke, eds. Berlin: Springer, 14–25.

Sawtooth Software (2013), *The MaxDiff System Technical Paper*. Orem, UT: Sawtooth Software, Inc.

Shaparenko, Benyah, Oezguer Cetin, and Rukmini Iyer (2009), "Data-Driven Text Features for Sponsored Search Click Prediction," in *ADKDD '09 Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*. New York: Association for Computing Machinery, 46–54.

Tibshirani, Robert (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B. Methodological*, 58 (1), 267–88.

Vestergaard, Torben, and Kim Schroder (1985), *The Language of Advertising*. Oxford, UK: Basil Blackwell Ltd.

West, Mike (2003), "Bayesian Factor Regression Models in the "Large p, Small n" Paradigm," *Bayesian Statistics*, 7, 723–32.

Zou, Hui, and Trevor Hastie (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B. Methodological*, 67 (2), 301–20.