SELECTING HIGH-PERFORMANCE ADS IN DESIGN CONTESTS

ABSTRACT

In this study, we examine how advertisers in crowd-based design contests can select ad designs with good click-through performance. Design contests allow advertisers to solicit many potential ad designs from the crowd. However, the large number of design submissions increases the challenges in evaluating and selecting the best designs in the contests. To help us identify factors of click-through performance and develop an efficient method to evaluate a large number of designs, we conducted a design contest experiment and acquired 340 ad designs from the crowd. We then developed a novel way to determine the design distinctiveness of these ads. Our approach to measure distinctiveness is significantly more efficient than those using traditional pairwise or spatial arrangement methods. We then launched an advertising campaign to measure click-through performances of the ads. We found that design distinctiveness and design quality jointly and positively affected the number of click-through received. Ads that are distinctive and well-designed obtained more clicks than other ads in the campaign.

Keywords: Crowdsourcing, design contests, online advertising, banner ads, click-through, design distinctiveness, design quality

INTRODUCTION

Advertisers often use click-through performance as a measure of online advertising effectiveness. In this study, we look at how advertisers in crowd-based design contests can select ad designs that are likely to achieve good click-through performance. Design contests involve advertisers broadcasting their design requirements to a crowd. After receiving design submissions from designers in the crowd, advertisers would select one or a few designs in their respective contests. Only designers whose submissions are selected would receive compensations in the form of contest prizes.

It is not uncommon for advertisers to receive a large number of submissions in design contests. While receiving many alternative designs increases advertisers' options, it also raises the challenges in evaluating and selecting the "best" designs. For example, to objectively measure distinctiveness of individual designs, advertisers need to conduct pairwise comparisons for all submissions in the contests. Given a set of *n* designs, there are $\frac{n(n-1)}{2}$ pairs of designs to compare. When *n* is large, the number of pairwise comparison to make is non-trivial (e.g., Kornish and Ulrich, 2011).

There are two main objectives in this study. First, we examine how design distinctiveness and design quality affect click-through performances of ads. Second, we develop a novel method to efficiently and objectively measure design distinctiveness in crowd-based contests. To achieve these objectives, we acquired 340 banner ads through a design contest. We then developed an approach to evaluate the design distinctiveness of these ads, and conducted an advertising campaign to measure their click-through performances. The results from the advertising campaign show that the number of click-through received depended on the joint effect of design distinctiveness and design quality. Distinctive ads that are well designed received more clicks than other ads in the advertising campaign.

CLICK-THROUGH PERFORMANCE IN ADVERTISING CAMPAIGNS

The number of banner ads served to online users is substantial, with more than 1.1 trillion display ad impressions delivered to U.S. Internet users during the first quarter in 2011.[1] Advertisers are estimated to spend US$22 billion on online display advertisements by 2015, overshadowing the spending on search advertising.[2] Click-through performance is an important consideration for most advertisers. In a survey of advertising agencies, a majority of the agencies (86%) indicated that they always or frequently used click-through performance to measure effectiveness of banner advertising (Shen, 2002). The other frequently used measures by the agencies included conversion (72%), ad exposure (53%), brand awareness (35%), and ad awareness (35%). In contrast, only 4% of the agencies indicated that click-through performance was seldom or never used as a measure of advertising effectiveness – this figure is relatively low compared to those who said they seldom or never used conversion (10%), ad exposure (27%), brand awareness (31%), and ad awareness (41%) as advertising performance measures.

Despite the importance that advertisers place on click-through performance, the click-through rates (CTR) in most ad campaigns are relatively low (between .09% to .20% on average).[3] Achieving better click-through performance would mean higher returns on advertising expenditures. Ads that attract more clicks could lead to greater customer awareness and conversions (e.g., enquires or purchases) as viewers who click on the ads are likely to be interested in what the ads are promoting.

AD ACQUISITION, SELECTION, AND IMPLEMENTATION

Conducting online banner advertising campaigns typically involves ad acquisition, ad selection, and ad implementation. There are various ways through which advertisers can *acquire* potential ad designs. One acquisition approach is contract sourcing, where advertisers enter into contracts and work with individual designers or ad agencies. Another approach that is becoming increasingly popular is crowdsourcing, where advertisers launch contests on crowd-based contest platforms to solicit ad designs (Koh, 2013). Regardless of the design sourcing approaches, the

[1] http://www.comscore.com/Press_Events/Press_Releases/2011/5/U.S._Online_Display_Advertising_Market_Delivers_1.1_Trillion_Impressions_in_Q1_2011
[2] http://www.emarketer.com/PressRelease.aspx?R=1008432
[3] http://www.mediamind.com/sites/default/files/MediaMind_Global_Benchmark_Q4_2010.pdf

deliverable in this stage is a set of designs for the advertisers' considerations. Having a relatively large and diverse set of alternative designs in this stage is important, as the likelihood of identifying high-performance designs increases with the number and variety of design submissions (Girotra et al. 2010).

In the next stage, advertisers would *select* one or few designs from the pool of alternatives to use in their advertising campaigns. A common criterion in the selection process is design quality, where visually appealing and attractive designs are usually preferred. Although one would expect high quality ads to achieve better click-through performance, recent studies did not find significant direct relationships between design quality (e.g., raters' evaluations of visual appeal and tastefulness of the ads) and CTR (Dow et al., 2010; 2011). While high design quality might have helped online ads to differentiate themselves during the early days of the Internet, this factor alone may not lead to high click-through performance given the clutter of quality ads that consumers are now exposed to. Hence, high design quality could be an insufficient condition for superior ad performance in today's crowded online advertising space.

After selecting the designs to use, advertisers would *implement* the advertising campaigns. They have to consider various factors such as ad locations (e.g., directly on specific websites, and/or through advertising platforms such as Google Display Network) and fees structure (e.g., cost per click or impression). Advertisers can also optimize their campaign settings so as to improve ad performance. For example, Google provide options for advertisers to optimize their advertising campaigns for clicks or conversions.

JOINT EFFECT OF DESIGN DISTINCTIVENESS AND DESIGN QUALITY ON CLICK-THROUGH PERFORMANCE

The main research question in this study relates to the ad selection stage – what types of ad designs are likely to achieve better click-through performance? To make their ads standout in a crowded online advertising space, advertisers can use distinctive ad formats, such as animations (instead of static/non-animated ads), in their advertising campaigns. Compared to static ads, animated ads generally performed better in terms of click-through performance, ad recall, and attitude towards the ads and/or brands (Diao and Sundar, 2004; Li and Bukovac, 1999; Rosenkrans, 2009; Sundar and Kim 2005). At times, however, advertisers cannot use

certain ad formats because advertising platforms impose restrictions on permissible ads. For example, Facebook does not support animated or flash-based ads, and Google limits animation length in banners to 30 seconds.[4] As a result, advertisers compete on a relatively level field in terms of ads formats on these important advertising platforms. They need to identify other design characteristics that could differentiate their ads from the competition, holding ad format constant.

Extending the concept of distinctiveness, we believe advertisers should consider *design distinctiveness* when selecting ads for their advertising campaigns. Design distinctiveness is a design's contrastive value relative to other designs (Jacoby and Craik, 1979; Rosenkrans 2009): The more different a design is from others, the more distinctive that design is. Ads with distinctive designs evoke a sense of surprise and unexpectedness (Jackson and Messick, 1965), which help to overcome viewers' resistance to the ads (Kover, 1995). Unique ad designs draw more attention and promote viewers' exploration of the ads. For example, in an experiment that used eye-tracking technology, viewers paid more attention to ads that are more unique and distinct (Pieters et al., 2002). Hence distinctive ad designs are likely to be more effective in attracting viewers' attention and clicks.

Advertisers can thus use design contests as a platform to source for novel ad designs. Ad designs that are distinctive in the contests have a higher chance of being distinctive in the online advertising space. However, design distinctiveness, by itself, may not be enough attract clicks. For example, a distinctive ad that is poorly designed may turn viewers away. Hence we hypothesize that design distinctiveness and design quality jointly affect click through performance. Viewers are more likely to click on ads that are distinctive *and* well-designed.

> *H1: Design distinctiveness and design quality jointly and positively affect the number of click-through that an ad receives.*

METHODS

Overview. We conducted a design contest where participants ("designers") were asked to design banner ads to promote an online wedding photography directory ("aweddinglist.com").

[4] See http://www.facebook.com/help/?page=245316378826196 for Facebook's advertising guidelines, and http://support.google.com/adwordspolicy/bin/static.py?hl=en&topic=1310862&guide=1308145&page=guide.cs&answer=176108&rd=1 for Google's.

Designers were asked to submit ad designs that are attractive and would achieve high ad recognition performance and click-through rate. To make our experiment realistic to design contests, we did not compensate designers for participating in this study. Instead, designers who submitted the top three designs would receive between US$250 and US$600.

During the contest, designers were given a logo of the online directory, and ten photos that they could use in their ad designs (Figure 1).[5] These photos showed different wedding-related images such as the bride and/or groom (in various poses and different settings), wedding bouquet, and wedding gown. Due to legal and copyright concerns, designers must only use the photos that we provided in their designs. Designers could include ad copy, such as tagline and phrases, in their submissions. The ad dimensions must be 300 (width) x 250 (height) pixels, with file size less than 50kb in file size, consistent with the requirements on Google Display Network.

[Figure 1 here]

Designers had ten days to submit as many designs as they wanted through the platform. We did not reveal the number of participants who were taking part in this study or the number of designs that had been submitted so as to minimize the impact of competition on designers' efforts during the contest. We also did not reveal submitted designs during the contest so that designers could not strategize their designs based on their observations of the competition.

Subjects. We recruited participants from various online communities for graphic designers. 180 individuals responded to a survey that we conducted before the start of the contest. 105 of these individuals subsequently submitted at least one design during the contest. We received 385 ads at the end of the contest. 27 ads did not meet the width and/or height requirements. As we could not resize these ads without removing key elements in the designs, we excluded them from our sample. 18 other ads included photos that we did not provide and/or URL of other websites instead of the online directory that they were supposed to design the ads for. We did not use these designs as we planned to launch an advertising campaign to assess the designs' performance. Therefore, we have 340 usable ad designs from 99 participants in our sample (3.43 designs per designer on average).

[5] A wedding photographer, who was blinded to the experiment, granted us the permission to use these photos. This photographer took all these photos during different weddings that she covered.

There were 50 females and 49 males in our sample. 71.7% of the designers had or were pursuing graphic design-related certificate or degree programs. The average designer had 8.4 years of design experience, and participated in 3.7 wedding-related graphic design projects in the last two years. Using the non-parametric Wilcoxon-Mann-Whitney test, we compared these statistics with those of the 81 individuals who were not in our sample. The two groups of individuals differed only in terms of design experience: on average, designers who were not in our sample had fewer years of design experience (mean = 6.8, std. dev. = 8.5) than those who were in our sample (mean = 8.4, std. dev. = 7.5) ($z = -2.309$, $p < .05$).

MEASURES

Design Distinctiveness. A traditional way to evaluate design distinctiveness is to ask raters to evaluate the creativity, novelty, and/or originality of a design (e.g., Aribarg et al., 2010; Dow et al., 2010; Heiser et al., 2008; Pieters et al., 2002). This approach assumes raters have perfect information of all available designs, and they can assess how a design is different from other designs. We took a different approach in this study, and developed a three-step method to measure design distinctiveness. First, we *codified* the attributes of individual designs in our contest. Second, we *compared* the difference between each pair of designs in terms of their attributes. Finally, we *calculated* the distinctiveness score of each design by averaging the design's pairwise distances with all the other designs in the contest.

Step 1 – Codify Design Attributes: The objective in this step is to describe attributes of each design. The attributes relate to the color scheme, photos, logo, and text in the design.

Color scheme. We extracted the color of each pixel in the designs so that we could quantify the color scheme used in each design. The color of each pixel can be expressed as an RGB triple that represents the amount of red (R), green (G), and blue (B) that the color has. The values for each RGB component range from 0 to 255. For example, black has rgb(0,0,0) whereas white has rgb(255,255,255). Using the RGB decimal values allows us to precisely describe the ad colors. For example, we could distinguish different shades of gray in a design, where some shades are more similar to black (e.g., rgb(24, 24, 24)) and others to white (e.g., rgb(168, 168, 168)). With the RGB decimal values for each pixel in a particular design, we calculated the proportion of every RGB triple in that design.

Photos, logo, and text. During the contest, designers could use different photos that we provided in the project brief for their work. They could also resize the logo, and use different amount of space for the ad copy. Two coders extracted information about these various elements in the designs. On a web-based interface, we displayed the designs individually and a set of "element-boxes" that represent the different elements that appear in the designs (Figure 2A). The coders first "activated" the respective element boxes by placing them over the corresponding elements (Figure 2B). (We created the element boxes with opacity of 80% in Adobe Photoshop and saved the files in transparent GIF format so the coders could see the actual elements on the designs underneath the boxes.) Next, coders adjusted the size of the element boxes to capture the area of the corresponding elements. Using the height and width of the boxes for the photos, logo, and text, we obtained the areas occupied by these elements in the designs. We also used the activated element-boxes to infer the *specific photos* and the *number of photos* in each design. The coders coded all the designs individually, and we averaged their scores for each element in the respective designs.

[Figures 2A and 2B here]

Step 2 – Compared Pairwise Differences in Design Attributes: By the end of Step 1, we had the attributes of each design in our sample. In Step 2, we would compare the designs so that we could calculate the distinctiveness of individual designs in the final step. We conducted the comparisons in two sub-steps. First, we compared the ads in terms of the various attributes (Step 2.1). Next, we used these attribute-level differences to estimate differences between designs at the design-level (Step 2.2).

Step 2.1 (Attribute-level comparison). Using the information from Step 1, we measured the differences in various design attributes between each pair of designs. First, we measured the difference in color schemes between two designs, *i* and *j*, by:

$$\sum_{x}\sum_{y} c_{x,i}c_{y,j}\sqrt{(r_{x,i}-r_{y,j})^2+(g_{x,i}-g_{y,j})^2+(b_{x,i}-b_{y,j})^2} \quad [1]$$

where $c_{x,i}$ is the proportion of color *x* in design *i*, and $r_{x,i}$, $g_{x,i}$, and $b_{x,i}$ are the respective decimal values for red, green, and blue of color *x* in *i*. (The same description applies to terms

with subscripts *y* and *j*.) The square-root term is the Euclidean distance between colors *x* in design *i* and color *y* in design *j*. Colors that are relative similar have a shorter distance between them. We weighted the Euclidean distance by the product of the proportions of the respective colors in each design; we assumed that colors that appear frequently in the designs have greater influence on the perceived dissimilarity between the designs. For each pair of designs, we compared every color in one design with all the colors in the other design. We summed all the weighted color comparisons between the designs to derive the difference in their color schemes.

The pairwise difference in terms of numbers of photos, size of logo, and size of text-space is expressed by:

$$\frac{|k_i - k_j|}{k_i + k_j} \quad [2]$$

where *k* is one of the above design attributes. We normalized the absolute difference between designs *i* and *j* in terms of *k* (numerator) by the sum of *k* in the two designs (denominator).[6]

We next measured the difference in terms of specific photos in the two designs by:

$$\frac{number\ of\ photos\ unique\ to\ i\ only + number\ of\ photos\ unique\ to\ j\ only}{number\ of\ distinct\ photos\ in\ i\ and\ j} \quad [3]$$

The value of this measure ranges between 0 (when designs *i* and *j* used the exact same photo images) and 1 (when the two designs have completely different photo images).

Two particular designs can include identical photos but they may still appear different because the sizes of the respective photos are not the same across the designs. Hence, we measured the difference in the sizes of specific photos that appear in each pair of designs:

$$\sum_h \frac{|area\ of\ photo\ h\ in\ i - area\ of\ photo\ h\ in\ j\ |}{2\ x\ 300\ x\ 250} \quad [4]$$

[6] Suppose there are two pairs of designs, designs A-B and C-D. In the first pair, the sizes of the logo in A and B are 1000 pixel^2 and 950 pixel^2, respectively. In the second pair, the sizes of the logo in C and D are 100 pixel^2 and 50 pixel^2, respectively. Although the difference in the logo sizes in both pairs is 50 pixel^2, this difference is likely to be perceived as larger in the second pair (C-D).

The numerator in the above equation is the difference in the sizes of a particular photo, *h*, in the two designs, and the denominator is the total area of the designs. We aggregated the ratio across all photos that appear in both designs.

Step 2.2 (Design-level comparison). With 340 designs in our sample, we needed to compare *57,630 pairs of designs.* To efficiently accomplish this, we built a model to estimate the dissimilarity between designs using the differences in design attributes from Step 2.1. We randomly chose 74 designs (21.8%) from our sample to use as a learning set for our model. In generating this sub-sample, we selected at most one design from each designer.

Five raters evaluated the distances (or dissimilarity) among designs in our learning set using the spatial arrangement method (SpAM) (Goldstone, 1994; Hout et al., 2012). Using SpAM, each rater arranged multiple stimuli simultaneously such that similar stimuli were placed closer to each other. This approach of collecting similarity/dissimilarity data is relatively fast and efficient. In a lab experiment to scale 25 to 27 stimuli, participants took about 5 minutes to complete the task using SpAM, and 25 to 30 minutes using traditional pairwise procedure (comparing the similarity of two items at a time using a Likert scale) (Hout et al., 2012). Moreover, Hout et al. (2012) found that the results using SpAM were comparable to those using pairwise procedures. In their study, the correlations between SpAM and pairwise procedure results ranged from .44 to .96.

We developed a web-based SpAM interface for raters to organize the 74 designs (2,701 pairs of designs) in our learning set (Figure 3). Because of the size and numbers of designs, we displayed six randomly selected designs on the webpage at a time. We scaled the designs to 180 x 150 pixels so that raters could work on the task without scrolling the webpage. All raters indicated that they could see all the designs clearly on their screens. We asked raters to arrange the designs on a 750 x 750 pixels white canvas, such that designs that were more similar were to be placed closer together.

[Figure 3 here]

The raters took about 31 seconds to arrange one set of six designs, and they each organized 245 sets on average (std. dev. = 72). Each time a rater completed organizing one set of

designs, our system would record the coordinates of the top-left corner of each design on the white canvas. Because the designs had the same dimensions, the distance between the top-left corners of two designs represented the distance between these designs. We normalized the distances between each pair of ads in every rater-set by the mean distance and standard deviation in the set. This normalization process helped to account for heterogeneity among raters[7]. Every pair of designs was evaluated by at least three raters. For each pair of designs, we averaged the normalized pairwise distances across all raters to derive its pairwise distance.

Next, we used a ten-fold stratified cross validation approach to build a model for estimating the pairwise distances in our full sample (Kohavi, 1995). We first split our SpAM data into ten mutually exclusive random subsets. Each subset contained the same pairwise distance distribution as in the full SpAM dataset. We then performed ten runs of our algorithm by regressing the pairwise distances on differences in design attributes. Appendix A shows the descriptive statistics and correlations of the variables in our model. The pairwise distance was obtained through the SpAM procedure, and the other variables were the differences in design attributes that we derived in Step 2.1.

In each cross-validation run, a different subset of data was used as a testing set, and the rest of the data was used as the training set. The results are relatively similar across the ten cross validation models (Columns 1 to 10 in Appendix B), and the normalized root mean square error (NRMSE) is .128. The differences in all the attributes, except that for the area of text in the designs, are positively and significantly related to the pairwise distances ($p < .001$). According to the averaged beta or standardized coefficients, differences in the specific photos used in the designs have the strongest effects on pairwise distances (beta = .315). This is followed by differences between the designs in terms of size of specific photos (beta = .232), color scheme (beta = .190), number of photos used (beta = .175), and logo size (beta = .107).

We estimated the distance between all the 57,630 pairs of designs in our sample using the averaged coefficients from the ten cross validation models (Column 11 in Appendix B). To further validate our model, we randomly selected five pair of designs within the 95% confidence

[7] Although we instructed the raters to use the entire canvas space that was provided, some raters might have used a more restricted space on the canvas than others. The unstandardized pairwise distances from these raters could therefore be systematically shorter.

interval at the 5th, 25th, 50th, 75th and 95th percentile in terms of the estimated pairwise distance. We also selected five pairs of designs with the highest estimated distance, and five pairs with the lowest estimated distance. We recruited users on Amazon Mechanical Turk (MTurk) to rate the similarity of these 35 pairs of designs on a 7-point scale ("not similar at all… extremely similar"). We reverse-coded and averaged the scores to obtain the dissimilarity scores. The MTurk scores and our model-estimated pairwise distances highly correlate (r = .91, p < .001).

Step 3 – Calculate Design Distinctiveness: Conceptually, a design is distinctive when it is, on average, relatively different from other designs. Thus, to calculate the distinctiveness of a particular design, we averaged that design's estimated pairwise distances with all other designs in the sample.

Advertising Campaign Performance. We conducted an advertising campaign on Google Display Network (GDN) to measure actual performances of the ads. GDN allowed a campaign to include multiple ad groups, and each group can have multiple ads. We randomly assigned the 340 ads into 5 groups (68 ads per group). GDN provided three alternatives for displaying ads: (i) optimize campaign for clicks, (ii) optimize campaign by conversion, and (iii) rotate the ads more evenly. We chose the option to evenly rotate the ads, as this removed a possible confound that the click-through performance was driven by GDN's algorithm. We also chose to place our ads on wedding-related websites to minimized the likelihood that our ads would appear in irrelevant websites.

We ran the campaign for 56 days, and tracked the number of impressions and clicks that each ad received. At the campaign level, we received 1,291,938 impressions and 2,054 clicks in total, achieving CTR of .16%. Each ad received between 3,520 and 7,790 impressions (mean = 3799.82, std. dev. = 414.40), and between zero and 25 clicks (mean = 6.05, std. dev. = 3.24). The CTR for individual ads ranged from 0% to .50% (mean = .16, std. dev. = .08).

The duration of our advertising campaign was relatively long compared to those in earlier studies (e.g., 12 to 15 days in Dow et al. (2010; 2011)). Furthermore, our campaign CTR of .16% is near the CTR in the industry and in other studies (e.g., .04% to .09% in Dow et al. (2010;

2011)). More importantly, based on the number of impressions received and the wedding-related websites that we selected to show our ads, each design had at least 3,520 targeted exposures. Thus, the performance data that we collected should be sufficient and appropriate for this study.

Perceived Design Quality and Expected Ad Performance. We invited six advertising industry professionals ("experts") to measure other design characteristics of the 340 ads. We randomly assigned between 170 and 228 ads to each expert. (One expert dropped out after evaluating 84 ads due to her work commitment.) The experts rated the designs individually and remotely through our experiment website.

When the experts logged on to the design evaluation website for the first time, we showed them the judging criteria. These criteria were consistent with those that we told designers in the experiment. First, the experts were to evaluate a design's *attractiveness* (or how visually appealing the design is) on a scale of 0 (not attractive at all) to 100 (extremely attractive). Second, to rate the *potential ad recognition performance* of the designs, they were to estimate the likelihood that a potential user would recognize a design one week after he or she had seen it for the first time. Third, the experts were to predict the *potential CTR* of the designs, between 0% to 5%, in an online campaign on GND. (We highlighted to the experts that the 0-5% CTR range was typical in online ad campaigns.)

Three experts evaluated each design. According to test-retest reliability checks, the experts were reasonably consistent in their evaluations over time despite the number of designs that were assigned to them. We thus averaged the experts' ratings of the respective criteria for each design. That is, the score of each criterion for a design is the average of the experts' evaluations.

We derived two measures of design characteristics based on the experts' evaluations. First, we averaged the attractiveness and expected ad recognition performance scores to compute the *perceived design quality* (alpha = .93). Second, we measured the *expected ad performance* by using the estimated potential CTR for the ads. This variable can account for factors such as the perceived effectiveness of the call to action in the ad designs.

Average cost per click. The wedding-related websites that we chose for the ad campaign differed in terms of their popularity and web traffic levels. As GDN did not provide details of the websites where the clicks occurred, we used the *averaged cost per click* for each ad during the campaign as a proxy of the quality of the websites on which the clicks occurred. We set a US$0.70 per click limit in the ad campaign; GDN charged us up to this amount for each click, depending on where the click took place. We assumed popular websites could command a higher cost per click.

RESULTS

We divided the number of impressions in our dataset by 1,000 to facilitate the reporting of estimated coefficients. We then mean-centered all the independent variables (Appendix C). Next, we estimated our model using fixed effects Poisson regression with dummy variables for individual designers (Cameron and Trivedi, 1998). We also included dummy variables to control for the groups that the ads were assigned to in the advertising campaign. Table 1 shows the results from standard Poisson regression.

[Table 1 here]

The estimated coefficients in Column 1 represent changes in logs of number of click-through for a one-unit change in the respective predictors, *ceteris paribus*. (To facilitate the interpretations of the effects, we showed the average marginal effects of individual predictors in Column 2.) The number of clicks is positively related to (i) number of impressions (in thousands) ($\beta = .445$, $p < .001$), (ii) average cost per click ($\beta = 1.240$, $p < .01$), and (iii) experts' estimations of the expected click-through rate ($\beta = .241$, $p < .05$). *We also found a positive interaction between design distinctiveness and design quality ($\beta = .054$, $p < .05$), supporting our hypothesis.* This result implies that the number of click-through depends on the joint effects of the two design characteristics, and distinctive and well-quality designed ads are more likely to achieve better click-through performance.

DISCUSSION

The range of CTR outcomes (between 0% and .5%) from our advertising campaign in Google Display Network showed that selecting the right (or wrong) ad designs in the contests

can significantly impact advertisers' return on advertising expenses. Hence advertisers need to be strategic in choosing the "best" designs during the ad selection process. In this regard, this paper makes two important contributions. First, we found that design distinctiveness and design quality jointly affect click-through performance. Second, we developed an efficient approach to objectively and efficiently measure distinctiveness of ad designs in the context of a design contests. Our results also address recent findings that show that design quality does not significantly predict click-through performances (Dow et al., 2010; 2011). Well-designed ads may be common and numerous in the Internet, and it is challenging for individual ads to distinguish themselves simply by their design quality. However, in this study, we found that design quality does positively affect click-through performance through its joint effects with design distinctiveness.

Implications for Advertisers. Advertisers compete for the share of viewers' attention and clicks in a crowded online advertising space. Design contests can serve as local competitions or trials for advertisers to select designs that give them the best shot as they compete in the global marketspace. From this perspective, advertisers should not approach designs contests simply as "the easiest and most affordable way to buy graphic designs"[8], or treat contests as events "where thousands of designers compete to create a design that [advertisers] love"[9]. Instead, contests are a channel for advertisers to explore design solution spaces and discover potentially novel design concepts through the crowd.

In other words, the best contests for ad designs may not be the ones that are easiest to run, most affordable, or give designs that advertisers would love. Advertisers should be strategic in structuring the contests and deciding which designs to acquire. The focus in this study is on the selection process – which design among all the submitted ones in a contest should an advertiser choose? This focus complements those in studies that examine other aspects such as contest compensations (e.g., Terwiesch and Xu, 2008) and designer characteristics (e.g., Boudreau et al., 2011; Jeppesen and Lakhani, 2010). Based on our findings, we recommend advertisers to place a greater emphasis on design distinctiveness when choosing ad designs. Previous studies also indicated other benefits of distinctive ads: such ads achieved higher recall of advertised claims,

[8] http://www.crowdspring.com/how-it-works/
[9] http://99designs.com/howitworks

better attitudes toward the ads and brands, and higher purchase intention compared to competing ads (Heiser et al., 2008; Keller, 1991). Therefore, advertisers ought to use distinctive ads to differentiate from competition and obtain better responses from users.

Implication for Design Contest Platforms. The large number of entries in design contests brings along a challenge in measuring distinctiveness: The number of pairwise comparisons that is needed to determine distinctiveness increases at a quadratic rate with the number of designs. Hence as more designs are received in the contest, trying to identify distinctive designs is like searching for a needle in the haystack. Furthermore, the rules on design contest platforms require advertisers to quickly decide on the winning designs within a certain number of days. Therefore, they cannot test the performance of individual design submissions before making their selection decisions. By assisting advertisers to efficiently select high-performance ads, contest platforms increase advertisers' returns on contest and marketing expenditures. Such value-adds can help platforms improve client satisfaction and retention. Advertisers who achieve good ad campaign performances using ads from a particular platform are likely to run subsequent contests on that platform.

In this study, we presented an efficient approach to measure distinctiveness in design contests. Our procedure requires human intervention only for codifying individual designs, while it uses an algorithm for the more resource intensive task of estimating pairwise differences between designs. We estimated that the time required to compare 340 designs using our approach (4 man-hours) is approximately one-seventeenth and one-eighth that for the traditional pairwise method and SpAM, respectively (see Appendix D). Hence this method is scalable and can handle a reasonably large number of design submissions in contests. Design contests platforms can consider implementing our approach to help advertisers identify distinctive designs among the received entries.

References

Aribarg, A., Pieters, R. and Wedel, M. “Raising the BAR: Bias Adjustment of Recognition Tests in Advertising,” *Journal of Marketing Research (47)*, 2010, pp. 387-400.

Boudreau, K. J.; Lacetera, N. & Lakhani, K. R. “Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis,” *Management Science (57)*, 2011, pp. 843-863.

Cameron, A. C. and Trivedi, P. K. *Regression Analysis of Count Data.* Cambridge University Press, 1998.

Diao, F. and Sundar, S. S. "Orienting Response and Memory for Web Advertisements: Exploring Effects of Pop-Up Windowand Animation," *Communication Research* (31:5), 2004, pp. 537-567.

Dow, S. P., Fortuna, J., Schwartz, D., Altringer, B., Schwartz, D. L. and Klemmer, S. R. "Prototyping Dynamics: Sharing Multiple Designs Improves Exploration, Group Rapport, and Results", *CHI 2011*, 2011.

Dow, S. P., Glassco, A., Kass, J., Schwarz, M., Schwartz, D. L. and Klemmer, S. R. "Parallel prototyping leads to better design results, more divergence, and increased self-efficacy," *ACM Trans. Comput.-Hum. Interact.* (17:4), 2010, pp. 18:1-18:24.

Girotra, K., Terwiesch, C. and Ulrich, K. T. "Idea Generation and the Quality of the Best Idea," *Management Science* (56:4), 2010, pp. 591-605.

Goldstone, R. "An efficient method for obtaining similarity data," *Behavior Research Methods* (26), 1994, pp. 381-386.

Heiser, R. S., Sierra, J. J. and Torres, I. M. "Creativity Via Cartoon Spokespeople in Print Ads," *Journal of Advertising* (37:4), 2008, pp. 75-84.

Hout, M. C., Goldinger, S. D. and Ferguson, R. W. "The Versatility of SpAM: A Fast, Efficient, Spatial Method of DataCollection for Multidimensional Scaling," *Journal of Experimental Psychology: General*, 2012, Forthcoming.

Jackson, P. W. and Messick, S. "The person, the product, and the response: conceptual problems in the assessment of creativity1," *Journal of Personality* (33:3), 1965, pp. 309-329.

Jacoby, L. L. and Craik, F. I. M. "Levels of Processing in Human Memory", *in* Cermak, L. S. and Craik, F. I. M., ed.,, Lawrence Erlbaum Associates, Hillsdale, NJ, 1975, pp. 1-21.

Jeppesen, L. B. and Lakhani, K. R. “Marginality and Problem-Solving Effectiveness in Broadcast Search,” *Organization Science (21)*, 2010, pp. 1016-1033 .

Keller, K. L. "Memory and Evaluation Effects in Competitive Advertising Environments," *Journal of Consumer Research* (17:4), 1991, pp. 463-476.

Koh, T. K. “Impact of Client-Provided Examples on Design Distinctiveness in Crowd-Based Design Contests,” *Working Paper*, 2013.

Kohavi, R. “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *Proceedings of the 14th international joint conference on Artificial intelligence (2)*, 1995, pp. 1137-1143.

Kornish, L. J. and Ulrich, K. T. "Opportunity Spaces in Innovation: Empirical Analysis of Large Samples of Ideas," *Management Science* (57:1), 2011, pp. 107-128.

Kover, A. J. "Copywriters' Implicit Theories of Communication: An Exploration," *Journal of Consumer Research* (21:4), 1995, pp. 596-611.

Li, H. and Bukovac, J. L. "Cognitive Impact of Banner Ad Characteristics: An Experimental Study," *Journalism & Mass Communication Quarterly* (76:2), 1999, pp. 341-353.

Pieters, R., Warlop, L. and Wedel, M. "Breaking Through the Clutter: Benefits of Advertisement Originality and Familiarity for Brand Attention and Memory," *Management Science* (48:6), 2002, pp. 765-781.

Rosenkrans, G. "The Creativeness and Effectiveness of Online Interactive Rich Media Advertising," *Journal of Interactive Advertising* (9:2), 2009, pp. 18-31.

Sundar, S. S. and Kim, J. "Interactivity and Persuasion: Influencing Attitudes with Information and Involvement," *Journal of Interactive Advertising* (5:2), 2005, pp. 5-18.

Terwiesch, C. and Xu, Y. "Innovation Contests, Open Innovation, and Multiagent Problem Solving," *Management Science* (54:9), 2008, pp. 1529-1543.

Figure 1: Logo and Photos in Project Brief

Figures 2A and 2B: Codifying Design Attribute

Figure 2A

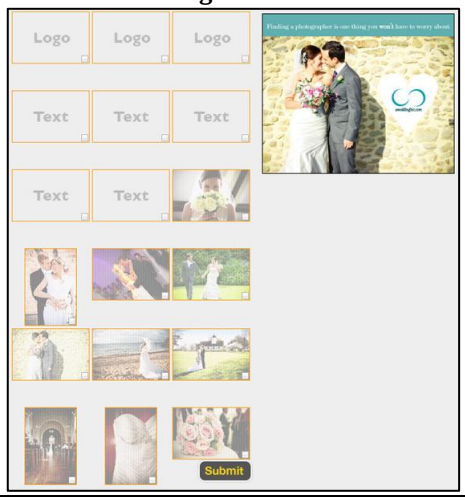

Figure 2B

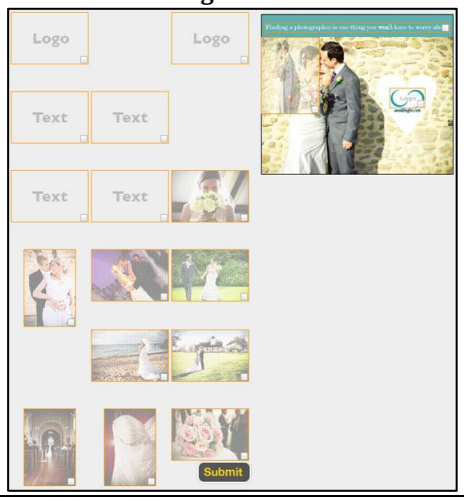

Figure 2A The coder saw the ad design (right) and the element boxes (left).

Figure 2B The coder dragged and placed the element onto the design. Next, he resized the boxes to fit the sizes of the corresponding elements on the design.

Figure 3: Spatial Arrangement Method (SpAM) Interface

Six ads were randomly selected in each round. Raters would drag the ads onto the white canvas, and arrange the ads such that similar ads were placed closer together.

Table 1: Poisson Estimation of Click-Through Performance

	[1]	[2]
DV: Number of Click-Through	**Coef.**	**Average Marginal Effects**
Constant	2.100*** (0.328)	-
Impressions (in thousands)	0.445*** (0.077)	2.691*** (0.466)
Average Cost Per Click	1.240** (0.422)	7.493** (2.558)
Ad Group 1	0.069 (0.079)	0.425 (0.5)
Ad Group 2	0.141+ (0.08)	0.892+ (0.525)
Ad Group 3	0.121 (0.081)	0.762 (0.529)
Ad Group 4	0.325*** (0.08)	2.184*** (0.595)
Expected CTR	0.241* (0.113)	1.454* (0.681)
Perceived Design Quality	-0.006+ (0.004)	-0.038+ (0.023)
Design Distinctiveness	0.504 (0.341)	3.043 (2.061)
Design Distinctiveness x Perceived Design Quality	0.054* (0.024)	0.324* (0.147)
χ^2	252.30***	

N = 340. Individual designer fixed effects are not shown.
Standard errors in parentheses.
$^{+}p < .10$ $^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

Appendix A Descriptive Statistics and Correlation Matrix for Pairwise Distance Estimation

		1	2	3	4	5	6	7
1	Pairwise Distance	1.000						
2	Difference in color schemes	0.164	1.000					
3	Difference in specific photos used	0.512	0.079	1.000				
4	Differences in number of photos	0.164	-0.105	0.071	1.000			
5	Difference in size of photos used	0.417	-0.156	0.722	-0.076	1.000		
6	Differences in size of logo	0.126	0.027	0.021	0.038	0.004	1.000	
7	Differences in size of text area	0.002	-0.093	-0.019	0.070	0.064	-0.034	1.000
	Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Std. Dev.	0.560	36.525	0.247	0.308	0.242	0.312	0.266
	Min	-2.033	-57.982	-0.870	-0.305	-0.709	-0.533	-0.409
	Max	1.532	201.504	0.130	0.695	0.409	0.467	0.591
N = 2701. All independent variables were mean-centered.								

Appendix B Cross Validation Results for Pairwise Distance Estimation

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
DV: Pairwise Distance	**Cross Validation Models**										**Average Coefficient**
Constant	0.004	-0.003	0.000	0.000	-0.002	0.000	0.000	-0.001	0.001	0.000	0.000
Difference in color schemes	0.003*** (0.187)	0.003*** (0.191)	0.003*** (0.183)	0.003*** (0.200)	0.003*** (0.189)	0.003*** (0.188)	0.003*** (0.185)	0.003*** (0.189)	0.003*** (0.204)	0.003*** (0.189)	0.003 (0.190)
Difference in specific photos used	0.738*** (0.328)	0.707*** (0.308)	0.709*** (0.312)	0.693*** (0.304)	0.738*** (0.325)	0.729*** (0.320)	0.712*** (0.314)	0.704*** (0.313)	0.700*** (0.308)	0.709*** (0.314)	0.714 (0.315)
Difference in numbers of photos	0.315*** (0.173)	0.345*** (0.189)	0.301*** (0.166)	0.322*** (0.177)	0.312*** (0.172)	0.323*** (0.179)	0.314*** (0.173)	0.320*** (0.177)	0.311*** (0.171)	0.308*** (0.169)	0.317 (0.175)
Difference in sizes of photos used	0.521*** (0.226)	0.562*** (0.240)	0.526*** (0.227)	0.565*** (0.245)	0.497*** (0.215)	0.522*** (0.226)	0.525*** (0.227)	0.535*** (0.232)	0.574*** (0.248)	0.537*** (0.233)	0.536 (0.232)
Difference in sizes of logo	0.197*** (0.109)	0.182*** (0.101)	0.192*** (0.108)	0.187*** (0.105)	0.199*** (0.111)	0.189*** (0.105)	0.193*** (0.107)	0.195*** (0.108)	0.182*** (0.102)	0.207*** (0.115)	0.192 (0.107)
Difference in sizes of text area	0.004 (0.002)	-0.003 -(0.002)	0.010 (0.005)	-0.016 -(0.008)	0.006 (0.003)	0.020 (0.009)	0.032 (0.015)	0.003 (0.002)	-0.007 -(0.003)	0.000 (0.000)	0.005 (0.002)
Number of Design Pairs	2431	2431	2431	2431	2431	2431	2431	2431	2431	2430	—
Beta coefficients in parentheses.											

Appendix C Descriptive Statistics and Correlation Matrix for Click-Through Performance Analyses

		1	2	3	4	5	6	7	8	9	10
1	Number of Click-Through	1.000									
2	Impressions (in thousands)	0.330	1.000								
3	Average Cost Per Click	0.053	-0.056	1.000							
4	Ad Group 1	0.007	0.176	0.037	1.000						
5	Ad Group 2	0.044	0.114	-0.050	-0.250	1.000					
6	Ad Group 3	-0.034	-0.102	0.002	-0.250	-0.250	1.000				
7	Ad Group 4	0.066	-0.267	0.079	-0.250	-0.250	-0.250	1.000			
8	Expected CTR	-0.021	-0.026	-0.059	-0.057	-0.051	0.112	-0.098	1.000		
9	Perceived Design Quality	-0.079	-0.019	-0.114	-0.063	0.034	0.042	-0.094	0.756	1.000	
10	Design Distinctiveness	0.032	0.046	0.049	0.020	-0.009	0.076	-0.123	-0.059	-0.048	1.000
	Mean	6.041	0.000	0.000	0.200	0.200	0.200	0.200	0.000	0.000	0.000
	Std. Dev.	3.237	0.414	0.069	0.401	0.401	0.401	0.401	0.384	11.666	0.097
	Min	0.000	-0.280	-0.543	0.000	0.000	0.000	0.000	-0.580	-20.863	-0.191
	Max	25.000	3.990	0.157	1.000	1.000	1.000	1.000	1.632	41.693	0.486
N = 340. All independent continuous variables were mean-centered.											

Appendix D

We estimated the time needed to compare distances (differences) among designs using three approaches: (i) traditional pairwise comparison method, (ii) Spatial Arrangement Method (SpAM), and (iii) model-based estimation method (which we developed in this study). We assumed:

1. Raters/coders do not experience fatigue during the tasks, and their reliability does not deteriorate over time, and
2. Raters/coders are perfectly substitutable and so there is no "downtime" in the tasks.

Pairwise comparison: In Hout et al. (2012), participants took 25-30 minutes to scale 25-27 stimuli using the traditional pairwise comparison method. Using the least conservative estimate, raters would need 25 minutes to compare 351 pairs of designs (27 x 26 x .5). Hence, it would have taken 68.4 man-hours to compare all the 340 designs (57,630 pairs) in our sample.

SpAM: On average, each rater in our study took 31 seconds to arrange 6 designs per set, and thus computed 15 pairwise distances (6 x 5 x .5) in each round. To obtain all pairwise distances for 340 designs, it should take approximately 33.1 man-hours using SpAM.

Model-based estimation: In our study, each coder codified 340 designs in three hours. It took approximately another hour to run the algorithm to compare characteristics among designs on our web-servers (2x Quad-core Xeon processors, 8GB RAM). Hence the total time needed using this approach is 4 man-hours. This requirement is approximately one-seventeenth and one-eighth that of the traditional pairwise method and SpAM, respectively.

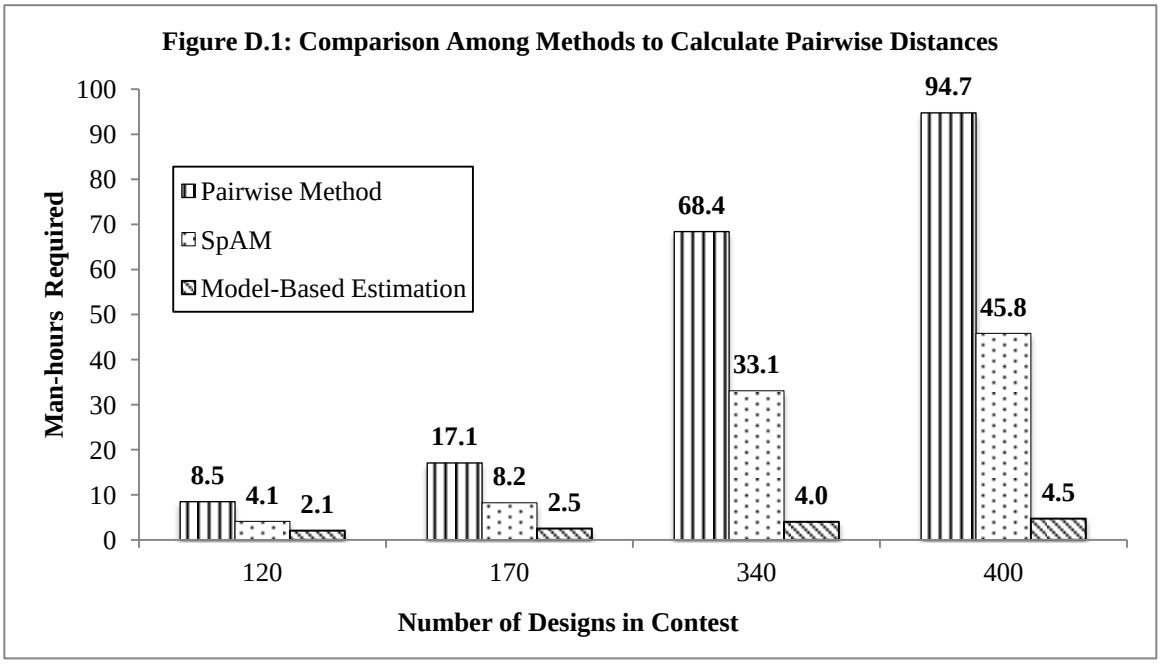

Figure D.1 shows the estimated time needed to calculate pairwise distances for different numbers of designs in a contest.[10] In all cases, the pairwise method requires the most time, whereas the model-based estimation approach requires the least. The amount of time saved using model-based approach increases exponentially with the numbers of designs. Under realistic conditions, the assumption that there is no rater fatigue is less likely to hold for traditional pairwise method and SpAM. As the number of designs increases, the cognitive load is much higher when using these two methods because raters have to make more comparisons among designs. In contrast, the estimation approach that we developed in this study only requires coders to codify individual design, which is less demanding cognitively. Hence, the estimated time saved using model-based approach should be conservative.

[10] We assume a one man-hour requirement to run the algorithm in our model-based approach for all cases.