Getting Users' Attention in Web Apps in Likable, Minimally Annoying Ways

Dan Tasse
Carnegie Mellon University
Pittsburgh, PA, USA
dantasse@cmu.edu

Anupriya Ankolekar
Hewlett Packard Labs
Palo Alto, CA, USA
anupriya.ankolekar@hpe.com

Joshua Hailpern
HP Labs and HP ArcSight
Sunnyvale, CA, USA
joshua.hailpern@hp.com

ABSTRACT
Web applications often need to present the user new information in the context of their current activity. Designers rely on a range of UI elements and visual techniques to present the new content to users, such as pop-ups, message icons, and marquees. Web designers need to select which technique to use depending on the centrality of the information and how quickly they need a reaction. However, designers often rely on intuition and anecdotes rather than empirical evidence to drive their decision-making as to which presentation technique to use. This work represents an attempt to quantify these presentation style decisions. We present a large (*n=1505*) user study that compares 15 visual attention-grabbing techniques with respect to reaction time, noticeability, annoyance, likability, and recall. We suggest glowing shadows and message icons with badges, as well as more possibilities for future work.

Author Keywords
User interface design; attention; alerts; notifications

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Web applications often need to present the user new information in the context of their current activity: social media sites must alert users to incoming chat messages, e-commerce websites need to notify shoppers about changes to items in their cart, and advertisements need users to click or even just look at them.

When choosing how to attract a user's attention, designers face a wide array of choices. With little empirical evidence about their relative effectiveness, they often make these important design decisions based on gut feelings, aesthetics, established norms, or perceived business constraints. Thus, designers risk making sub-optimal design decisions to the detriment of users and their own end goals.

For example, pop-ups are widely used, perhaps because designers believe they are effective at capturing users' attention, but they end up annoying and repelling users [2]. Further, using the same visual widget for high value and low value information runs the risk of users ignoring that widget entirely, as users do with large banner ads [5]. In the worst case, users may be turned off and abandon their task or the website entirely. Subtler notifications may be less annoying, but they may also be less effective at getting attention. As a result, end users needs' and designers' end goals may still not be in line.

This work quantitatively explores a sample of the large spectrum of attention grabber presentations. While there exists substantial research into *when* best to interrupt users [1, 11, 20], we have few quantitative studies about *how* to visually present users with new information. This paper addresses this question with a set of empirically grounded design guidelines for effective and appealing visual techniques to get users' attention in web pages.

The primary contributions of this work are the results of our experiment to quantify Attention Grabbers by varying presentation parameters. Through a large (n = 1505) quantitative study on Amazon Mechanical Turk, we examined 15 different Attention Grabbers in terms of their effectiveness in capturing users' attention, their likability and recall of information. From our results, we suggest presentation styles and directions for future work to determine the optimal means to get a user's attention.

RELATED WORK
Human attention is a finite resource. As Kahneman's Capacity Theory [22] posits, the only way to get a user's attention is to divert it from something else. The amount of information in our world has exploded, but our attention capabilities have stayed constant [32]. This has led to "economies of attention" [18], as people have realized the monetary and social value associated with being able to attract attention.

This work seeks to optimize a small part of the web app experience. This work is in a similar vein to [14], which tested a number of different interrupters against each other; ours expands it and situates a similar experiment in a

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05···$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858541

modern web application. Other similar work includes [15, 16], which explored optimization of one UI component: how to visually present progress bars to seem the quickest. However, given the breadth of the field of interruption research, it is important to explain how our work relates, and why we made the choices that we did.

Interruption Research & Timing

While the focus of this paper is squarely on the visual method used to grab users' attention, this work complements prior research studying why interruptions happen [8, 9]; the effects of interruptions on performance, resumption speed, and long-term stress [3, 7, 23]; and when to schedule interruptions [1, 11, 20]. This study answers an orthogonal question, "*How* should the system get the user's attention, assuming it must do so right now?" Ideally, the results of this work and future studies would complement the results of the existing interruption research.

Types of Attention Grabbers

The work that is most related to this paper involves previous attempts to figure out better and worse ways to get the user's attention. Designers often get users' attention with dynamic widgets because users often ignore static banners and text [5, 6, 29]. Some work has investigated pop-ups [2], moving icons [4], and animations [17]. We hope to continue in the vein of quantitatively comparing different techniques, as in [4], while expanding the study widely and testing in modern web apps.

To ground our work and guide our choice of attention grabbers, we look to research in notification systems. While our work, designing one element of a web app, differs from designing an entire system, work by McCrickard *et al* [26, 27] provides a useful framework. They distinguish between three axes: Interruption, Reaction, and Comprehension, as shown in Figure 1. Alerts that require interruption want the user to shift their attention to the alert, those that need reaction want the user to do something (regardless of whether they shift attention), and those that need comprehension want the user to store information and relate it to existing knowledge. Every notification or alert needs some combination of these three. We aim to investigate subjective preference of web apps at all points in this space.

SCOPE AND DEFINITIONS

The primary contributions of this work are the results of our experiment to quantify attention grabber methods by varying presentation parameters. We use the term "**Attention Grabber**" **(AG)** to refer any user interface element that tries to get the user to attend to it. This is used instead of "notification" in order to also include unrelated visual elements like advertisements.

The primary study of this work applies to web applications, as this was our sole testing platform. We use the term **Web Application** to include static web pages as well, because they can be seen as a special case of web application. We expect our results to generalize to single-window desktop applications as well, but we do not make any claims because we only tested web apps.

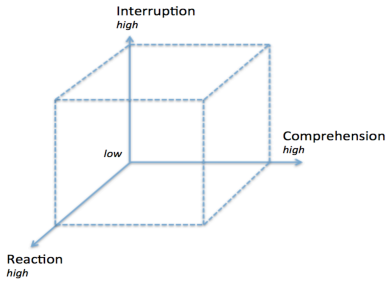

Figure 1 - The IRC framework created by McCrickard et al.

Outside of the scope of this work are system-level effects (e.g. Mac OSX Dock), cross application AGs (e.g. mail notifications when playing a game), mobile/ubiquitous applications, or auditory AGs. While some results of our work may apply more broadly, the other AG contexts bring an entirely new set of challenges and options due to the control and pervasiveness of the OS and smaller screen of the mobile device. Likewise, a full examination of people's reactions to different utilities of content is outside the scope of this paper.

EXPERIMENT DESIGN

In order to quantify varying visual presentations of AGs, we had to present different AGs to different people within a comparable context. To do this, we had participants play a game during which an AG would appear. AG style was varied between participants (independent variable). Dependent measures included participant reaction time, recall of AG content, and responses to survey questions. Our experiment employed a between subjects design, with each participant only seeing one AG during their session. The remainder of this section expands on our study design.

Experimental Context: Set Game

Participants played a variation of the game Set; in which players are given 12 cards (**a board**) with various symbols (diamond, oval, and squiggle), colors (red, purple, and green), shadings (open, filled, or striped), and symbol counts (1, 2, or 3). Participants had to find **Sets** of three cards that fulfill certain criteria: for each attribute (symbol, color, shading, and count), the three cards must all be the same or all differ. We made two variations from the original game: first, after each set is found, a new random board is displayed, thus refreshing the entire game environment, and preventing users from bringing any information from the previous "round" forward. Second, every board that is shown to users has exactly 4 possible sets within it (ensuring no board has an implicit advantage over another).

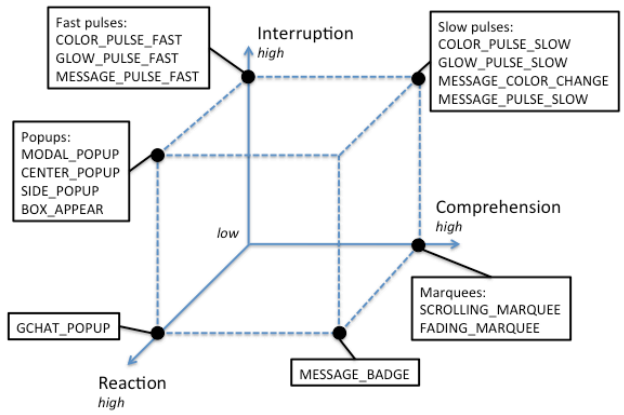

Figure 2 - The 15 attention grabbers used in this study, arranged in the IRC framework of McCrickard et al.

Set was chosen as an experimental context for three reasons. First, it is engaging and requires a lot of concentration (like many tasks users might be doing in a web application). Previous studies involving interruptions have also used interactive games, so we felt it was a reasonable context [13]. Second, participants could be remunerated based on their performance. This ensured participants had a "stake" in the outcome, unlike performing a task in a fictional context (e.g. booking a fake flight). Third, it is fun, so it was easy to recruit participants.

Motivation

Crowd workers are much more motivated when their pay is performance dependent [25]. Therefore, we created motivation through performance-based remuneration. Participants were paid $0.30 for completing the task, plus $0.15 for each Set found. They found a mean of 7.36 Sets per person (median = 7), resulting in a mean payment of $1.40 each. Thus the majority of a participant's remuneration came from performance, motivating them to play the game. As the game took a mean 14.05 (*SD* = 6.19) minutes per person (median = 12.63), this resulted in an hourly wage of $6.02.

Procedure

Participants were recruited from Amazon Mechanical Turk, an online marketplace for small tasks that require human intelligence. Upon accepting our task, participants would come to our site, go through interactive Set instructions, play a "training round" until they found two sets (demonstrating they understood the rules and were ready to play), and then play the "real" game for five minutes. The sets found in the five-minute round determined the bonus pay. Halfway through the five-minute game, at 2:30, an AG would appear with new information. At the conclusion of the game, participants would answer a questionnaire.

Independent Variable: Attention Grabbers

In this section, we describe all the AGs tested in this study, with Figure 2 as a reference for their theoretical IRC framing and Figure 3 as a reference for placement and visual style. In total, 15 AGs were tested plus an additional control group where no AG was displayed. We chose these 15 AGs with respect to the IRC framework, trying to have at least one that covered most possible combinations of interruption, reaction, and comprehension. The AGs we include are as follows: (we used the names in capital letters throughout the experiment to avoid confusion)

- Low interruption, reaction, and comprehension: this would be noise; no AG would try to do this, so we ignored it.
- High interruption: fast pulses. We included a box that pulsed orange quickly (COLOR_PULSE_FAST), a box that had a glowing shadow that pulsed quickly (GLOW_PULSE_FAST), and a message icon that pulsed quickly (MESSAGE_PULSE_FAST). These pulses (all with a period of 0.5 seconds) indicated that something was happening, but not that reaction or comprehension was necessary.
- High comprehension: marquees. We included a scrolling ticker of text (SCROLLING_MARQUEE) as it was McCrickard et al.'s example for this category [27], as well as a marquee that faded in and out (FADING_MARQUEE) instead of moving because of evidence that continuous motion can be distracting [23].
- High reaction: GCHAT_POPUP. This box appears as the word "Message" in a box in the lower right corner of the screen, expanding into a message box when the user clicks it. This design, inspired by the Google Chat (now Hangouts) indicator, was selected because it acted as an indicator and therefore invited reaction, without inviting much interruption or comprehension of what is being reacted to.
- High interruption and reaction: Pop-ups. Despite research showing their ineffectiveness [2], they are still widely used, and definitely cause interruption and reaction. We implemented four varieties: MODAL_POPUP (which prevented any interaction with the page behind it until it was dismissed), CENTER_POPUP (which also appeared in the center of the screen, but did not prevent interaction), SIDE_POPUP (which appeared on the right center side of the screen), and BOX_APPEAR (which appeared inside the page on the right side, instead of as a pop-up appearing "over" it.)
- High interruption and comprehension: Slow pulses. These appear closer to the "animation in place" that McCrickard *et al*. discuss, which could help users to

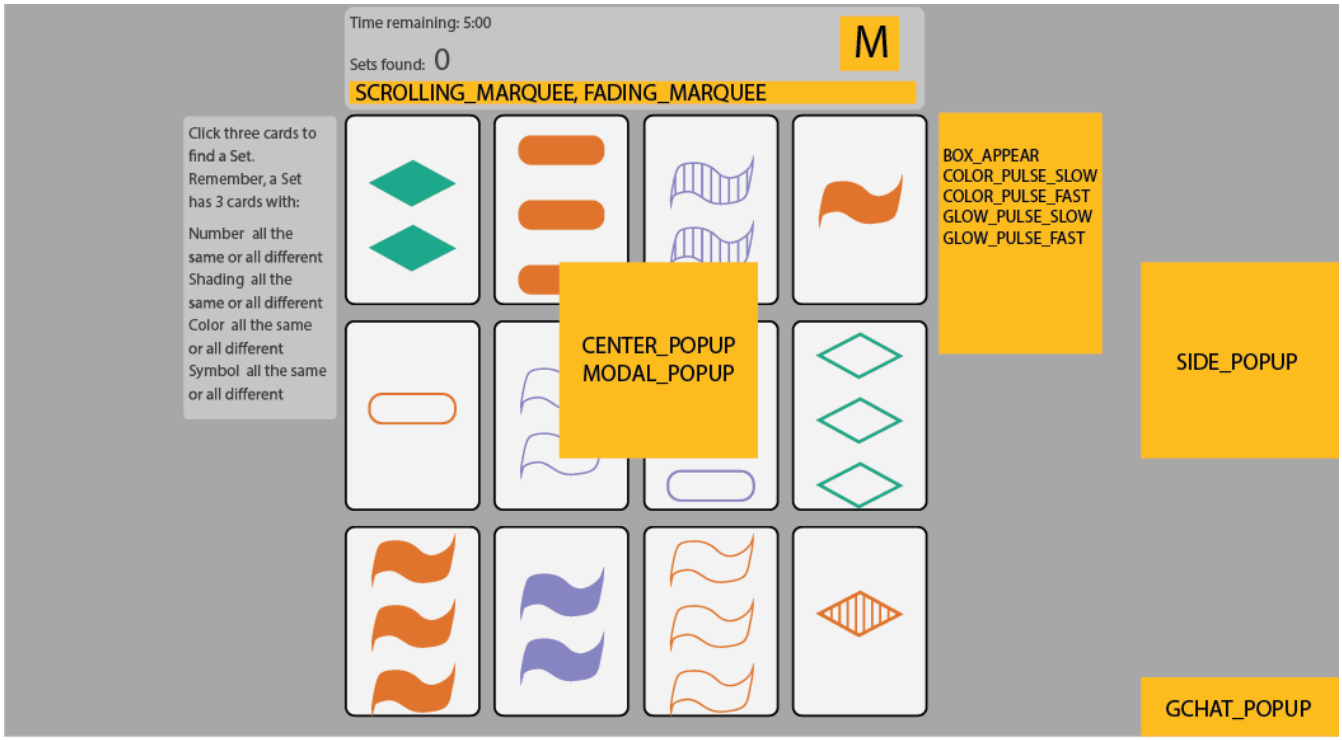

Figure 3a - Attention Grabber Layout Placement. The "MESSAGE" attention grabbers all appeared in the box marked "M".

Figure 3b - Attention Grabber Visual Styles

Figure 3 - Attention Grabber visual design and placement

slowly understand information without inducing reaction.

- High comprehension and reaction: A message icon with a badge (MESSAGE_BADGE). For this AG, a message icon was present throughout the study, and at one point a small number "1" appeared over it, as if to say "there is one message here." This AG has become a popular means of attracting attention on sites such as Facebook, Github, and Reddit. We also included a variant called MESSAGE_COLOR_CHANGE, in which the envelope turns orange instead of adding a badge.
- High interruption, comprehension, and reaction: This would be a critical activity monitor; few websites attempt to target all three of these at the same time.

Our mappings between attention grabbers and IRC goals began theoretically, but our evidence from the study later showed that our AGs indeed spanned a wide range of interruption, reaction, and comprehension. As it would be impossible to study all possible notifications, we considered this sample spanning the IRC range as a useful starting point.

	Comp.1 (noticeability)	Comp.2 (annoyance)	Comp.3 (likability)
As soon as the (attention grabber) appeared, I saw it immediately.	**0.91**	0.18	0.05
I noticed a (attention grabber).	**0.92**	0.18	0.05
The (attention grabber) grabbed my attention.	**0.89**	0.25	0.13
The (attention grabber) interrupted my thoughts.	0.28	**0.85**	-0.06
The (attention grabber) distracted me from playing Set.	0.35	**0.83**	-0.08
The (attention grabber) was annoying.	0.05	**0.80**	-0.28
The (attention grabber) was aesthetically pleasing.	0.08	-0.13	**0.86**
I liked the (attention grabber).	0.08	-0.29	**0.84**
I wish other web sites would use a similar (attention grabber) when they need to get my attention.	0.03	0.00	**0.81**

Table 1. Factor loadings on each principal component. Loadings over 0.4 are shown in bold.

Dependent Variables: Performance, Reaction Time, Survey, and Recall

As a measure of performance, we recorded how many Sets each participant found. For some of the AGs, which asked the user to click on them, we recorded the time it took

After they played the game, we asked participants three pages of survey questions – deeply grounded in prior literature. The first page contained a six-question NASA Task Load Index (TLX) scale [28], to assess perceived cognitive load of the entire game. Answers were on a 7-point Likert scale, from "Very Low" (1) to "Very High" (7). TLX questions were presented verbatim, and therefore applied to the whole task (e.g. "How successful were you in accomplishing what you were asked to do?") The second page contained questions about the game as a whole, to see if the one AG affected their entire experience. Some of these were taken from the EGameFlow scale, which was designed to measure enjoyment of e-learning games [12]. We included 11 of the 42 questions[1]; we did not include the full EGameFlow scale because many of the questions focused on e-learning games. The third page contained nine questions about the AG itself, which are described in the Results section. We used questions inspired by the survey in [2], but we had to adjust them slightly to address more than just pop ups. Answers on pages 2 and 3 were on a 7-point Likert scale, from Strongly Disagree (1) to Strongly Agree (7). We included these because we wanted to directly investigate how effective each AG was at getting users' attention, how annoying each AG was, and how much users liked or disliked the AGs.

We hoped to test recall, so we included an icon in each AG. After playing the game Set, each participant was asked via free text what the image was in the AG. In order to reduce ambiguity, a picture of an AG was included above the question with a "?" in the place where the image was located. For the first 1000 participants, the icon was a heart; for the last 920, it was a star. We realize that this is a different task than most lab-based recall tasks, because the icon is unrelated to their task and we never ask them to remember it, but we think this will make our task better reflect the real world. Many real-world attention-grabbing widgets, like ads, explicitly want a user to remember or pay attention to something unrelated to the current task. Remembering a star or a heart icon is analogous to noticing a product that is being advertised.

We also included an "attention check" math problem (e.g. "what is three plus one?") on each page to make sure that they were paying attention. If they failed these, we deleted their data, did not pay them, and reposted the task for another individual to complete.

Controls

All participants played the same game, for the same length of time, and saw an attention grabber at the same time for the same duration. Like [14], we counterbalanced subjects between three different messages, including one "high-utility", one "medium-utility", and one "low-utility" as their pertained to the game. To confirm as a control, our post-hoc analyses found no main or interaction effects.

Study Power

To determine the number of participants to recruit, we performed an a priori power calculation. As the majority of our outcome variables were 7-point Likert scale survey questions, we powered our study to detect small changes of effect size 0.1; this roughly corresponds to 14 groups rating something 4/7 and two groups rating it 5/7. Assuming that the variance within each group would be 1, we calculated power for an ANOVA, and got $n = 118$. Rounding up for convenience, we aimed to recruit 120 users in each group.

[1] Specifically, questions C5, C6, C7, C8, A7, I1, I2, I3, I5, G2, and F3.

Recruitment

Participants were recruited from Amazon Mechanical Turk. Participants were randomly assigned an AG and message type. This ensured that workers who found HITs quickly, or workers in certain time zones, would not cluster in certain conditions. We used two batches in order to expedite study execution, recruiting 1000 in the first batch and 920 in the second batch. 415 people participated in both batches, and later analysis showed significant differences in responses between repeaters and first-timers, so we excluded all 415 second runs from our data. These 415 were mostly evenly distributed between the latter 8 conditions (no attention grabber, both marquee conditions, all four message conditions, and the Gchat-style pop-up); all 8 of these groups still had between 61 and 81 first-time participants. While removing duplicates affects study power, the power loss is not critical. For completeness, we ran the same tests with all 1920 participants and our findings did not differ from what we report here, so we just report results on 1505 participants for simplicity.

Statistical Methods

Given that Likert Scale questions are not continuous data, and it cannot be assumed that our responses are normally distributed, we opted for the more conservative non-parametric tests when analyzing our Likert data. We emphasize that this is more conservative; any significant results that would be found with parametric tests will also be found with non-parametric tests. Thus a Wilcoxon rank sum was used to perform pair-wise comparisons (instead of a Student's T-Test), and the Kruskal-Wallis one-way analysis of variance (instead of a parametric ANOVA) comparing more than two groups of independent data[2]. If significance was found, we used Tukey's post hoc HSD to compare all pairs of groups.

We analyzed the following two questions to ensure that our manipulation of message type worked: "The information on the (AG) was related to the Set game" and "The information on the (AG) was necessary for the Set game." We compared participants who received the "low-utility" message with those who received the "medium-utility" message, and those who received the "low-utility" message with those who received the "high-utility" message.

We then analyzed the rest of the survey questions (TLX and the 17 other questions), first with the Kruskal-Wallis, then if significance was found, with pair-wise tests.

We also investigated the interruption lag, defined as the amount of time between when the AG first appeared and when it was dismissed (if applicable).

We noted whether the participant had correctly remembered the icon present in the AG. We marked their answer correct if they included the word "heart" or "love" for the first 1000, or "star" or "asterisk" (or simply typed the asterisk character) for the last 920. We manually double-checked the answers of those who failed this string matching.

Principal Component Analysis

In order to better distill our survey responses into higher-level concepts, we conducted a PCA on the 9 survey items, following [10], with orthogonal rotation (varimax). The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis with KMO = .8, and all KMO values for individual items were >.70, well above the limit of .5. Bartlett's test of sphericity, $\chi 2(36) = 8025$, $p < .00001$, shows that correlations between items are sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Two components had eigenvalues over Kaiser's criterion of 1, and one had eigenvalue 0.98. After analyzing the scree plot, we determined that three components were justified. These three components explained 79% of the variance. **Table 1** shows the factor loadings after rotation. A residual analysis showed that no further components were needed. The questions that loaded on each component suggest that the components represent noticeability, annoyance, and likability, respectively. Cronbach's alpha scores for each component were .92, .84, and .80, respectively, suggesting high reliability for each component. We took the mean of each participant's scores on each question within each component to find that participant's overall score for that component.

For each of the three principal components, we used the Kruskal-Wallis to compare responses under different AGs, finding significant results for each. We then used Tukey's post-hoc HSD to find where significant differences occurred. The results are shown in each graph.

PILOT TESTING

In order to confirm that the difficulty level of the game was appropriate, and to see if the two-Set training round was long enough, we piloted with 45 of our coworkers and friends. We plotted each Set our pilots found against the time it was found. We found that 42 of the 45 people found sets at a linear rate (with Pearson's r >.9). Therefore, we concluded that, for most participants, there was little learning occurring after the two set training round. (Furthermore, the number of Sets found was only one of many outcomes that we measured.)

If players received an unlucky board that was very difficult, they might become "stuck" and be unable to continue playing. Further, in our pilot data, we found that it took players less time to find Sets when there were more Sets on the board. To avoid unnecessary variation between participants, and to give the game an intermediate difficulty

[2] While performing multiple comparisons may suggest statistical adjustment to a more conservative value (i.e., Bonferroni correction), we choose to show our levels of significance following [31], to provide a more transparent view of our findings.

level, we ensured that every new board of 12 cards would contain exactly four possible Sets.

We also ran a few small rounds of pilot testing on Mechanical Turk, in order to fine tune parameters there. Based on this testing, we allowed workers who had completed at least 1000 HITs and had 98% accepted tasks.

STUDY RESULTS

The study was completed in 17 days: 8 days for the first 1000 participants and 7 days for the next 920 (with 2 days between the two rounds). During the study, we rejected 89 participants, based on the "attention check" survey questions, for a 4.4% rejection rate. We republished their rejected tasks so other participants could complete them. As mentioned in previously, we excluded data from 415 participants for repeating the task, in order to avoid potential confounds. We were left with data from 1505 participants.

We found significant differences in interruption lag between groups and in participants' preferences based on our three survey principal components of noticeability, annoyance, and likability. We found no significant differences between groups based on recall, NASA TLX or EGameFlow questions.

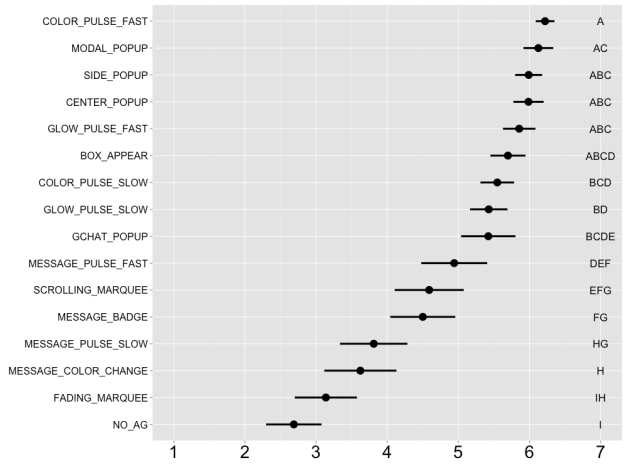

Figure 4 - Noticeability by attention grabber type, on a Likert scale from 1 (Strongly Disagree) to 7 (Strongly Agree). In this and in Figure 5 and Figure 7, error bars represent 95% confidence intervals. Two conditions that do not share a letter on the right side of the graph are significantly different by Tukey's post hoc HSD, $p < .05$

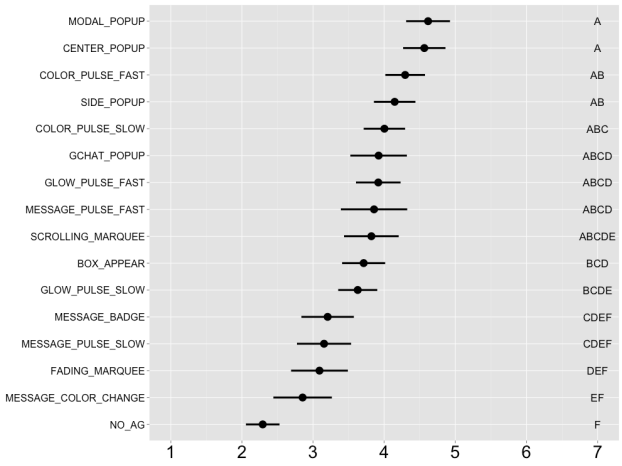

Figure 5 - Annoyance by attention grabber type. AGs are sorted differently in this, Figure 4, and Figure 7, in order to more easily show differences. The X axis runs from 1 (Strongly Disagree) to 7 (Strongly Agree), as in Figure 4.

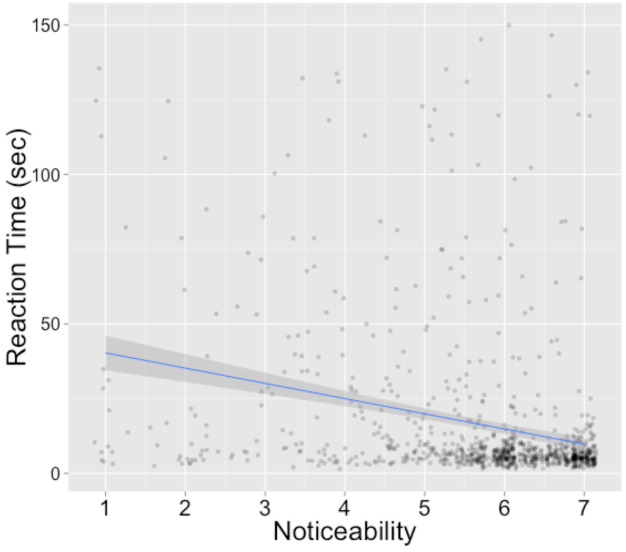

Figure 6 - Noticeability by reaction time. Noticeability is the mean of three answers. One point represents one user. Points are slightly jittered to avoid overplotting.

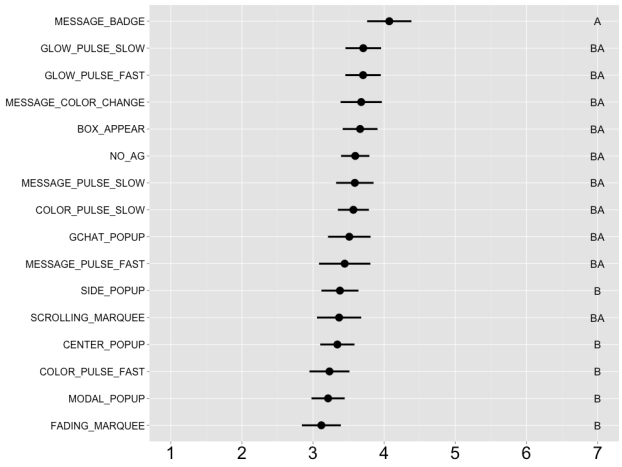

Figure 7 - Likability by attention grabber type. The X axis runs from 1 (Strongly Disagree) to 7 (Strongly Agree), as in Figure 4.

Component Analysis: Noticeability and Interruption Lag

In our Principal Components Analysis, the first principal component contained the following questions:

- **I noticed a (AG).**
- **As soon as the (AG) appeared, I saw it immediately.**
- **The (AG) grabbed my attention.**

As these all relate to how well the AG gets the user's attention, we decided to call this component "noticeability." Results are shown in Figure 4.

In addition, the interruption lag (or reaction time) statistics correlated with people's answers to these questions. A linear regression showed a correlation coefficient of -.27 ($p < 10^{-15}$) (Figure 6), which further strengthens our claim that the answers to the survey questions reflect the noticeability of the AGs.

Based on these results, it appears that the most noticeable AGs were the box with quickly pulsing color and the pop-ups, followed by boxes with pulsing shadows. The marquee and message options were the least noticeable.

Component Analysis: Annoyance

The following questions loaded on the second component:

- **The (AG) was annoying.**
- **The (AG) distracted me from playing Set.**
- **The (AG) interrupted my thoughts.**

We called this construct "annoyance", as these three characteristics were all facets of how annoying or distracting the AG is. Results are shown in Figure 5.

Results for this were similar to the Noticeability results, with the exception that pop-ups were more annoying than others, compared to how noticeable they were. Surprisingly, the AGs that required interaction were not necessarily the most annoying (e.g. COLOR_PULSE_FAST).

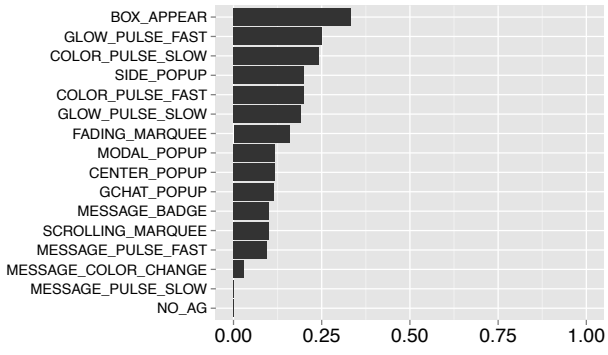

Figure 8 - Recall: Percent of participants who remembered the icon correctly

Component Analysis: Likability

The third component contained the following questions:

- **The (AG) was aesthetically pleasing.**
- **I liked the (AG).**
- **I wish that other sites would use a similar (AG) when they needed to get my attention.**

We called this component "likability". Results are shown in Figure 7. The message icon with a badge proved remarkably likable, while other message and "glow pulse" AGs seemed likable as well. Pop-up and marquee AGs were seen as the least likable. However, most of the differences are not significant, besides the message icon with a badge being more likable than some of the pop-ups.

Recall

We also examined whether people remembered the icon that was present along with the text. These results are shown in Figure 8. We found that boxes appearing, boxes with pulsing colors, and marquees resulted in the most correct answers, and pop-ups and message icon AGs led to the lowest recall. However, we present this finding with some caveats. Surprisingly, very few people answered the recall question correctly: 33% at best. Also, the Message Icon conditions may have scored poorly because a lot of participants answered "Envelope", indicating that they might not have understood which icon we were asking about. Third, it may have been more appropriate to ask about the message than about the icon, because that was the item that mattered to the participants. Because of these shortcomings, we chose not to speculate too much about the results of this measure. Nevertheless, these results do provide preliminary evidence that some AGs, such as pop-ups, do not cause users to pay attention to or remember them, perhaps because they quickly dismiss them.

Performance, TLX, Game Questions and Message Types

There were no significant differences by AG type for number of Sets found or TLX scores. Twelve of the 17 game-related questions returned significant differences between message types, but the effect sizes were so small that we do not report those either. These "significant" but low effect size results are most likely an effect of an overpowered study. In addition, most of the survey questions about the game did not show any significant results between AG types. Four of them did, but a Tukey's post-hoc HSD test revealed almost no pairs with significant differences.

Relationships Between Factors

There is a correlation between noticeability and annoyance ($r = 0.87$), but no significant correlation between either and likability. We also found no significant correlation between any of these three factors and recall.

Control Verification: Message Type

As mentioned previously, we provided different messages on the AG. Like [14], we thought that matching utility of messages might impact how positively people perceived our AGs, so we similarly offered a high-utility, medium-utility, and low-utility message. We measured that our manipulation here was effective. For both of our check questions, "The information on the (AG) was related to the Set game" and "The information on the (AG) was necessary for the Set game", Wilcoxon Rank-Sum tests between message types showed statistically significant results, $p < 10^{-15}$. Therefore, we are confident that our "medium-utility" condition was indeed related to the task and our "high-utility" condition was necessary. However, we found no significant differences on any of our dependent variables based on the message type or interaction effects, so we feel confident ignoring it, and treating it as a control.

DISCUSSION

Before discussing concrete suggestions, we hope to point out a few high-level results from our study. In the introduction to this paper, we alluded to the question of whether noticeable AGs must also be annoying. McCrickard's framework suggests this as well: if something is very interruptive, it is annoying almost by definition. This was supported by our data: noticeability and annoyance highly correlated.

However, in this study we hoped to get at a more subtle point: people's subjective reactions to these interruptors. Even the terms "interruptive" and "annoying" have different valences: interruption can be positive or negative, while annoyance is almost always negative. Similarly, we found some preliminary differences in likability, and while we may lack a sensitive enough scale to pick up on small differences, it is clear that some (like the message icon with a badge) are more appreciated by users than others.

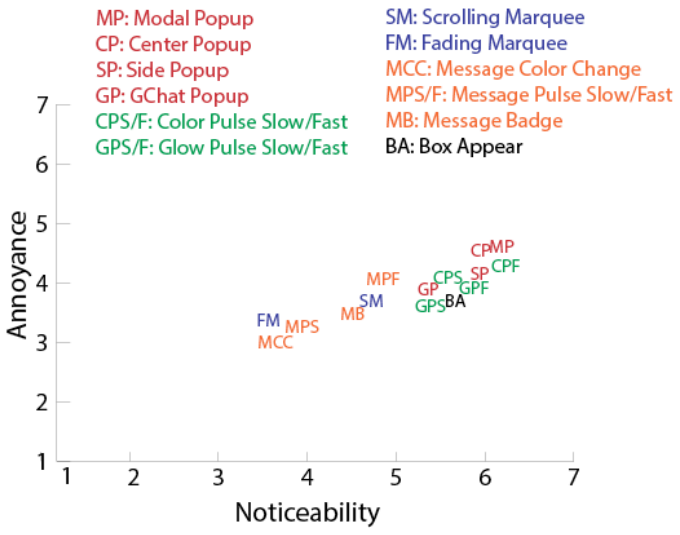

Figure 11 - Noticeability by annoyance. One point represents mean values for all users who saw that AG.

We also found no correlation between recall and noticeability, annoyance, or likability of an AG. Therefore, we conclude that further research is needed to understand what makes AGs memorable.

Methodologically, we found this study a fruitful way to study a lot of different AGs, with a lot of participants, very quickly. We hope that other researchers will follow this approach to further learn more about specific UI features.

Design Suggestions

Following McCrickard *et al*, we note that all notification or attention grabbing circumstances require a different combination of interruption, reaction, and communication. Based on our study, we have identified a preliminary recommendation for each corner of the IRC cube in Figure 11. These are of course only starting points, and individual designers must choose the right AG for their application.

In addition, we saw a few more salient points throughout the study, which we have turned into three main suggestions for web interface designers.

DR1. For immediate attention, use pulsing shadows

The two AGs containing pulsing shadows (glow pulse slow and fast) scored well on all measures. A fast pulsing shadow is statistically equally noticeable as a pop-up, but much more likable and less annoying (approaching statistically significantly). A slow pulsing shadow, while slightly less noticeable, is statistically as likable as, and less annoying than, any other AG. They were both acceptably memorable, as well. No other AGs scored as consistently well on all of our scales. This surprised us, as we have not seen pulsing shadows widely used. We propose these as potentially useful, and underused, design methods for getting attention in websites.

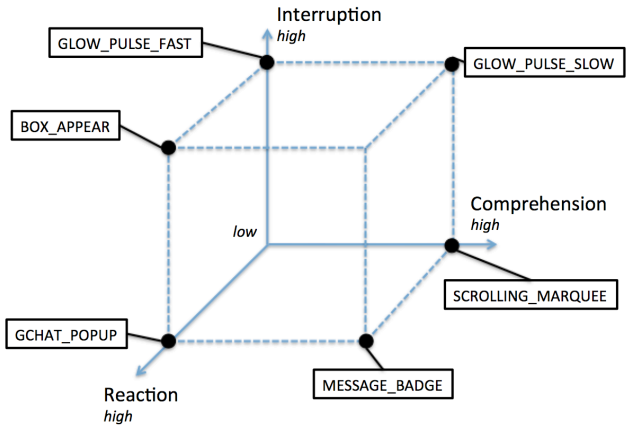

Figure 10 - Attention grabber recommendations based on different interruption, reaction, and comprehension needs. Note that these are only a starting point and they must be customized to individual applications.

DR2. For less-critical info, use an icon with a badge

A message box with a badge was the most likable AG, statistically more so than half of the AGs. It was also almost the least annoying. However, it does suffer in noticeability. Therefore, we recommend it for use cases where it is not critical that the user attend to the information immediately. These might correspond to the "low interruption" situations in the IRC model [26].

DR3. If something must pop up, make sure it integrates well with the page

Pop-ups provide quick reactions, but users rate them as statistically significantly more annoying than many AGs, and they score near the bottom in likability as well. A box simply appearing near their field of view was statistically equally noticeable, but liked more. Perhaps this is because it is integrated within the page, and therefore perceived as part of the page, instead of being a separate element that floats above the page. Another possibility is that pop-ups have become associated with useless ads or malware.

Other Observations and Recommendations

Marquees, both scrolling and fading, should be avoided. They scored poorly in almost all measures: participants found them not very likable or noticeable. Perhaps this has become obvious; they are much less prevalent than they used to be in the early days of the web. Also, previous research has shown issues with continuous movement and other features of marquees [23]. However, many news sites still use them, undoubtedly due to their success in television news shows. While viewers may like tickers across the bottom of the screen (among other enhancements) [21], this study provides evidence that they may not have the same preferences in web-based settings. This disparity might occur because the web may be a more attention-demanding medium than TV.

When designing a pulsing screen element, faster pulsing makes it more noticeable but less likable and more annoying. As a result, and informed by the failure of the web's blink tag, we advise exploring pulsing elements, especially fast pulses, with caution. Further work should be done to determine the ideal rate of pulsing.

FUTURE WORK & LIMITATIONS

One of the most important limitations of this work is the reliance on survey-based measures. Indeed, more controlled work, with more objective performance-based measures, would always be welcome. However, similar studies have been done in labs in the past (e.g. [14]) and so for this study we hoped to trade off the direct control possible in the lab for the ability to recruit thousands of users online. We see our work and lab-based work as complementary. Our approach could be a first step before undertaking the increased effort, cost, time and recruitment of a lab study, allowing researchers to quantitatively understand what dimensions to focus on in the future.

We studied 15 attention grabbers, so there may very well be others that perform better than anything tried. In addition, there are some parameters that could be further explored. For example, we studied only two speeds of pulsing icons and boxes, and only one location of each AG. As others [6, 29] have found, location of an element can play a role in how memorable it is and how likely it is to be ignored.

In addition, our participants only saw an AG once. Maybe after seeing one of them multiple times, they would get used to it and its noticeability, annoyance, or likability would change. Relatedly, due to fashions in application design, people's preferences may change over the years. As more websites start to use the same AGs, they may become common in people's minds and therefore less noticeable. Because of these changes over time, this paper serves as both an update to the work of McCrickard *et al*, while offering a study framework that can be replicated in the future.

Relatedly, future participants could do a more natural task, like looking up locations on a map or writing an email. The game Set offered a controllable testing environment, but of course attention requirements of real-world applications differ widely.

Finally, it would be quite natural to extend this study to other platforms, as well. Our results should work well on desktop or web applications, but due to the ways mobile devices integrate with people's lives, AGs for mobile devices may be quite different.

CONCLUSION

We presented a series of recommendations for application designers, based on data from a 1505-person study where participants played a game while a user interface element tried to get their attention. Based on their survey answers, reaction times, and recall, we identified UI elements with glowing shadows as the most likeable and effective way to get user attention. Icons with badges are a good alternative for less-critical information. We also found that users prefer dynamic visual elements that blend in with the surrounding content instead of pop-ups. Using these recommendations, designers can create user interfaces that are likely to be more useful, usable and appealing to users. We consider data-driven studies to improve user interfaces a promising avenue for research and encourage future work in this area.

REFERENCES

1. Piotr D. Adamczyk and Brian P. Bailey. If Not Now, When?: The Effects of Interruption at Different Moments Within Task Execution. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, (2004), 271–278.
2. G. Susanne Bahr and Richard A. Ford. How and why pop-ups don't work: Pop-up prompted eye movements, user affect and decision making. Computers in Human Behavior 27, 2 (2011), 776–783.
3. Brian P. Bailey and Joseph A. Konstan. On the need for attention-aware systems: Measuring effects of interruption on

task performance, error rate, and affective state. Computers in Human Behavior 22, 4 (2006), 685–708.

4. Lyn Bartram, Colin Ware, and Tom Calvert. Moving Icons: Detection And Distraction. IFIP TC.13 International Conference on Human-Computer Interaction (INTERACT), (2001), 157–166.
5. Jan Panero Benway and David M. Lane. Banner Blindness: Web Searches Often Miss "Obvious" Links. Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting, (1998), 463–467.
6. Moira Burke, Anthony Hornof, Erik Nilsen, and Nicholas Gorman. High-Cost Banner Blindness: Ads Increase Perceived Workload, Hinder Visual Search, and Are Forgotten. ACM Transactions on Computer-Human Interaction. 12, 4 (2005), 423–445.
7. Edward Cutrell, Mary Czerwinski, and Eric Horvitz. Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance. INTERACT, (2001), 263–269.
8. Mary Czerwinski, Eric Horvitz, and Susan Wilhite. A diary study of task switching and interruptions. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, (2004), 175–182.
9. Laura Dabbish, Gloria Mark, and Victor González. Why do I keep interrupting myself?: environment, habit and self-interruption. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, (2011).
10. Andy Field, Jeremy Miles, and Zoë Field. Discovering Statistics Using R. Sage, Los Angeles, (2012), 749-811.
11. James Fogarty, Scott E. Hudson, Christopher G. Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C. Lee, and Jie Yang. (2005). Predicting Human Interruptibility with Sensors. ACM Transactions on Computer-Human Interaction, 12(1), 119–146.
12. Fong-Ling Fu, Rong-Chang Su, and Sheng-Chin Yu. EGameFlow: A scale to measure learners' enjoyment of e-learning games. Computers & Education 52, 1 (2009), 101–112.
13. Tony Gillie, and Donald Broadbent. What makes interruptions disruptive? A study of length, similarity, and complexity. Psychological Research 50, 4 (1989), 243–250.
14. Jennifer Gluck, Andrea Bunt, and Joanna McGrenere. (2007). Matching attentional draw with utility in interruption. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 41. http://doi.org/10.1145/1240624.1240631
15. Chris Harrison, Brian Amento, Stacey Kuznetsov, and Robert Bell. Rethinking the Progress Bar. ACM Symposium on User Interface Software and Technology, (2007), 115–118.
16. Chris Harrison, Zhiquan Yeo, and Scott E. Hudson. Faster Progress Bars: Manipulating Perceived Duration with Visual Augmentations. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, (2010), 1545–1548.
17. Weiyin Hong, James Y.L. Thong, and Kar Yan Tam. How Do Web Users Respond to Non-Banner-Ads Animation? The Effects of Task Type and User Experience. Journal of the American Society for Information Science and Technology 58, 10 (2007), 1467–1482.
18. Bernardo A. Huberman and Fang Wu. The Economics of Attention: Maximizing User Value in Information-Rich Environments. Advances in Complex Systems 11, 4 (2008), 487–496.
19. Ipeirotis, P. Demographics of Mechanical Turk. (2010).
20. Shamsi T. Iqbal and Brian P. Bailey. Investigating the Effectiveness of Mental Workload as a Predictor of Opportune Moments for Interruption. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, (2005), 1489–1492.
21. Sheree Josephson and Michael E. Holmes. Clutter or content? How on-screen enhancements affect how TV viewers scan and what they learn. Eye Tracking Research & Application Symposium, (2006), 155–162.
22. Daniel Kahneman. Attention and effort. Englewood Cliffs, NJ: Prentice-Hall, (1973).
23. Paul P. Maglio and Christopher S. Campbell. Tradeoffs in displaying peripheral information. Proceedings of the SIGCHI conference on Human factors in computing systems, 1 (2000), 241–248.
24. Gloria Mark, Daniela Gudith, and Ulrich Klocke. The cost of interrupted work: more speed and stress. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, (2008), 8–11.
25. Winter Mason and Duncan J. Watts. Financial Incentives and the "Performance of Crowds." HCOMP, (2009).
26. D. Scott McCrickard, C.M. Chewar, Jacob P. Somervell, and Ali Ndiwalana. A model for notification systems evaluation---assessing user goals for multitasking activity. ACM Transactions on Computer-Human Interaction. 10, 4 (2003), 312–338.
27. D. Scott McCrickard and C.M. Chewar. User Goals and Attention Costs. Communications of the ACM 46, 3 (2003), 67–72.
28. NASA Task Load Index (TLX). Accessed from http://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX.pdf
29. Justin W. Owens, Barbara S. Chaparro, and Evan M. Palmer. Text Advertising Blindness: The New Banner Blindness? Journal of Usability Studies 6, 3 (2011), 172–197.
30. Zachary Pousman and John Stasko. A taxonomy of ambient information systems: four patterns of design. Proceedings of the working conference on Advanced visual interfaces, (2006), 67–74.
31. David A. Savitz and Andrew F. Olshan. Multiple comparisons and related issues in the interpretation of epidemiologic data. American Journal of Epidemiology 142, 9 (1995), 904-908
32. Herbert A. Simon. Designing organizations for an information-rich world. Computers, Communication, and the Public Interest, The Johns Hopkins Press (1969).