# Modeling Triage Decision Making

**J. Isaiah Harbison (jiharb@umd.edu),**
**Alan Mishler (amishler@umd.edu),**
**Thomas Wallsten (tswallst@umd.edu)**
Center for Advanced Study of Language, University of Maryland
7005 52nd Avenue, College Park, MD 27642

## Abstract

With the ever increasing amount of information available, the ability to prioritize the most relevant items for full processing is increasingly necessary to maintain expertise in a domain. As a result, accurate triage decisions--initial decisions about the relevance of a given article, book, or talk in order to determine whether to pursue that information further--are very important. In the present paper, we present a model of triage decision making that includes both an information search component to determine reading strategy and a decision making component to make the final decision. We apply the model to human relevance ratings as well as binary decisions of relevance for a set of emails.

**Keywords:** information search; information foraging; decision making; triage decision making

## Introduction

When looking through Google Scholar results or the latest issue of your favorite journal, how do you decide if an article is relevant to your research? This relevance decision or triage decision requires at least two components: an information search strategy to guide search through the abstract or article, and a decision making process to determine when sufficient evidence has been collected to make the decision of relevance.

In the present paper, we introduce a model of triage decision making that combines an information search process with a decision making mechanism. We then test the model against relevance judgment data and binary decisions.

## Searching for a Topic

Duggan and Payne (2009) found evidence that when skimming, people read until the rate of information gain drops below a threshold. At that point they either move on to a new section (paragraph, chapter, etc.) or they stop skimming the document. This behavior matches the behavior prescribed by the marginal value theorem (MVT). According to the MVT in order to optimize the number of items found, search in a given location or patch should continue until the rate of return drops beneath the expected rate of return in the general search environment. For example, in an online literature search, the rate of return for the general environment would be calculated by combining the average expected rate of information gain from other articles in the patch (e.g., each article would be considered a patch in the current application), including the expected time taken to find and open those articles. This rate of information gain establishes the threshold that the rate of gain from the current article is compared against.

The calculation of expected rate of return from general search relative to the rate of return of the current document would be very complicated, making the MVT implausible as a descriptive stopping or switching rule. However, there is another relatively simple stopping rule, the incremental rule that is able to implement MVT. The incremental stopping rule compares the total amount of time spent in search (or total number of search attempts performed so far) to a stopping threshold in order to determine whether to terminate the search. The threshold is incremented each time a target item is found. For example, suppose the initial stopping threshold is two minutes and the increment is five seconds. If a search proceeds for two minutes without returning any target items, then the search will be terminated. Each time a target item is found, however, five seconds are added to the time allowed to pass before terminated. After one item is found, the stopping threshold becomes two minutes and five seconds (i.e., the total time allowed for search from beginning to end would be two minutes and five seconds if only one item was found). If ten target items were found, the stopping threshold would be two minutes and fifty seconds. This strategy makes stopping decisions a function of the rate of return from a patch, ensuring that search continues longer in rich patches and terminates earlier in poor patches.

Evidence for the use of the incremental stopping rule has been found both for external search (Hutchinson, Wilke, & Todd, 2008) and internal search (e.g., memory search; Harbison, Dougherty, Davelaar, & Fayyad, 2009). Our model applies the incremental stopping rule to triage decision made about text. Information search proceeds by reading through a document until either the document has been read, the incremental stopping rule indicates search should be terminated, or the relevance of the document can be decided upon. The mechanism for latter case will be described next.

## Identifying a Topic

Triage decision making is a matter of both information search and decision making. Search can be terminated as a function of the rate of return of information, but it can also be terminated based on a decision of relevance being reached if sufficient evidence has been accumulated. In the former case, once search is terminated a decision must still be reached regarding the item's relevance. Given these two possibilities, how is the decision ultimately made about whether a text is relevant to a given topic?

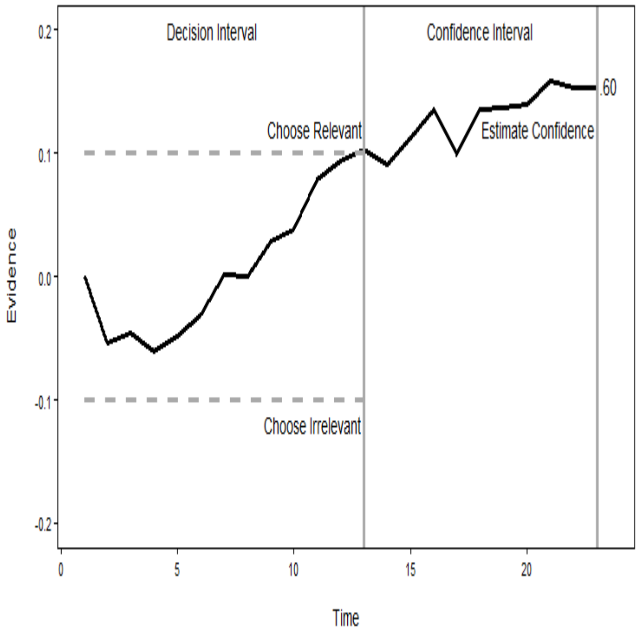

Figure 1. Schematic of the 2DSD decision making process, with a relevance judgment (first vertical line) and a confidence judgment (second vertical line).

We utilize the two-stage dynamic signal detection (2DSD) model of confidence and decision making (Pleskac & Busemeyer, 2010) to model the relevancy decision process. The 2DSD model conceptualizes the decision maker as moving between two decision thresholds. The model uses a random walk/diffusion process to account for movement between these two options, as shown in Figure 1.

In the case of triage decision making, the two decision thresholds are "relevant and "not relevant." During the decision process, the decision maker skims the material gathering evidence for and against an article's relevance. Each new piece of evidence moves the evidence state between two thresholds from its starting position. Evidence for relevance moves the evidence state toward the "Choose Relevant" threshold, and evidence against its relevance moves the evidence state towards the "Choose Irrelevant" threshold.

In previous uses of the 2DSD model, evidence was sampled from a random distribution that reflected the knowledge the individual had about the relative evidence for each of the two options. Here evidence is the information for and against each of the options. To the extent that evidence is clearly in favor one option, the probability of picking evidence in support of that option is increased, increasing the probability and speed of the dominant option being chosen (i.e., that option's threshold being passed). Once a threshold is crossed, a decision is made. If the decision maker is asked to give their confidence in their decision, this judgment is made after additional processing and additional consideration of evidence, as depicted on the right portion of Figure 1.

The difference between the present and previous applications of the 2DSD model (other than applying its application to triage decisions) is that the evidence is not sampled from a random distribution. Instead, the evidence is a function of the material being triaged. For the present modeling application, the materials are emails and the evidence is determined by the words being read. The evidential value of each word in our model consists of the point-wise mutual information (PMI, detailed below) between the word and the topic under consideration. The state of belief is determined by the running sum of the evidence values of the words read.

## Triage Decision Making Model

There are three components to the triage decision making model: a computational representation of topics, a search strategy, and a model of decision making. For the first component, we represented topics as probability distributions over words. This method is common both to the natural language processing of text for guiding internet search (Blei, 2012; Xu, Yang, & Li, 2009) and to psychological models attempting to reflect human understanding of topics (Steyvers & Griffiths, 2007). The core idea is that the topic under discussion influences the chance that any given word will occur. For example, words such as "computational", "model", and "theory" are more likely to occur in conversations about cognitive science than in a conversation about Philadelphia. Therefore, it would be reasonable for a decision maker who encounters these words while reading to judge the likelihood that the text is about cognitive science more highly than the likelihood that the text is about Philadelphia.

The second component is the skimming or search strategy. We tested the following three strategies:

1. The Complete Strategy. The searcher moves from the beginning to the end of the document, making a decision about the relevance of the document only after reading the entire text.
2. The Skim Beginning Strategy. The searcher moves from beginning to end, but search is terminated once the threshold of evidence is crossed or after the stopping rule indicates that search should be terminated without a decision being made.
3. The Skim Paragraph Strategy. The searcher makes decisions about staying within or leaving individual paragraphs of the document instead of the entire document. In this strategy, the decision is made either after the threshold is crossed or after the decision maker has searched and left each of the individual paragraphs.

The third component of the model is the decision process. We used a process similar to the random walk process from the 2DSD model, considering each word as evidence for or against a conclusion. As noted above, however, the decision process in our model is not truly a random walk. Since the evidence is represented as the PMI of each word to the topic, evidence is accrued through skimming the document through one of the search strategies. a systematic walk through the text instead of a random sampling of possible evidence.

## Model Details

### Information Ecology

To evaluate the model, we tested it against human judgments of topic relevance. Specifically, we applied the model to a set of emails drawn from the 2001 Topic Annotated Enron Email Data Set and compared the model's judgments to those of participants from a new experiment described below. This dataset consists of emails that have been hand-labeled with topics by a trained annotator, with one topic label per email (Berry & Browne, 2007). The topics used in our experiment were: California Analysis and Daily Business. California Analysis emails included executive summaries and analyses about the company's affairs in California. The Daily Business emails are about buying and trading shares on the stock market, setting up meetings, and confirming meetings, as well as general announcements from human resources.

### Topic Representation

We create the representations of each topic by calculating the frequency distributions of the words within each set of emails for that topic and converting these into probability distributions. The evidence in the model consists of the pointwise mutual information (PMI) between the topics and the term, calculated as follows:

$$PMI(term, topic) = log\frac{p(term,topic)}{p(term)p(topic)}.$$

Participants in the experiment were asked to read the emails one at a time and indicate whether each email was relevant to the target topic (California Analysis). Half the emails came from the target topic, and the other half came from the distractor topic (Daily Business). The distribution of PMIs for words in these two topics are shown in Figure 2. These distributions are approximately normal, with the target mean just above zero (.427) and a distractor mean just below zero (-.025).

Examining the two distributions of PMIs in more detail, we calculated the mean PMI of words as a function of the sentence and paragraph in which they occurred. Many of the emails had at least five paragraphs of five sentences each, so we looked at the mean PMI for the first through fifth sentence of the first through fifth paragraph. These means are shown in Figure 3. Across paragraphs there is little variation in mean sentence PMIs. Likewise, the variation in mean PMI of sentence within paragraph does not show a clear pattern.

Given the overlap in target and distractor email PMI distributions, it is not clear how well the stimuli can be used to identify topic emails. To get an impression of how the accumulation of evidence would work for emails from both topics, we plotted the cumulative evidence (PMI) of both types of emails. The running sum of the PMI between the words in the email and the target topic is shown in Figure 4. As the figure shows, overall the PMI for the distractor emails is lower than for the target emails, suggesting that even though the average PMI for target emails and the target topic is just above 0, the summation of the evidence provided by the emails can be used to distinguish between target and distractor emails.

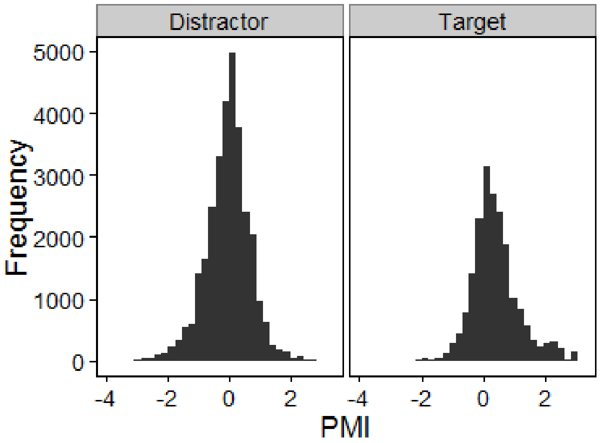

Figure 2. Distribution of pointwise mutual information values (PMI) for words within each topic.

### Skimming Process

We tested three ways of moving through the text to make a triage decision: Complete, Skim Beginning, and Skim Paragraph. All three methods start with the first sentence of the first paragraph of the email. The Complete (or full read) Strategy continues reading to the end of the document, accumulating all of the evidence provided by the words along the way but only making a decision at the end. This strategy is unique in that it does not allow for early search termination for any reason.

The Skim Beginning Strategy continues reading until the evidence collected reaches a stopping threshold ($\theta_S$) or when the incremental failure rule prescribes and end to the search (i.e., when a threshold, $\theta_F$, number of words fail to impact the current state of the evidence). Likewise, for the Skim Paragraph Strategy, search is terminated after a decision threshold is reached. The difference between the two skimming strategies is that the Skim Paragraph Strategy will not terminate search if the information gain in the current paragraph decreases below threshold but instead moves to the beginning of the next paragraph in the email. The goal of movement between paragraphs is to allow for more effective search, with the assumption that the most informative words are likely to be in the beginning of paragraphs. Note that the Skim Paragraph Strategy will terminate search once a decision of relevance is reached or once it leaves the final paragraph.

### Decision Process

Decisions are made in the model by a running sum of the evidence for and against relevance. The original application of the 2DSD followed a random walk/diffusion process (Pleskac & Busemeyer, 2010), shown in Figure 1, in which evidence is accumulated over time that moves the model closer or further away from the thresholds for determining between options ($\theta_D$ and $-\theta_D$, respectively). The random walk is instantiated by randomly sampling from a distribution of evidence values that move the current state of evidence either up or down. The distributions are different for target items than for distractor items. Drawing from the target distribution, on average, provides movement towards the "relevant" threshold and drawing form the distractor's distribution, on average, provides movement towards the "irrelevant" threshold. Once a threshold is crossed, the decision is made about the item's relevance. If participants are required to give a confidence or probability estimate, they do so after some amount of additional processing time in which it is assumed that more evidence is considered (either through additional reading or by reflecting on what has been processed so far).

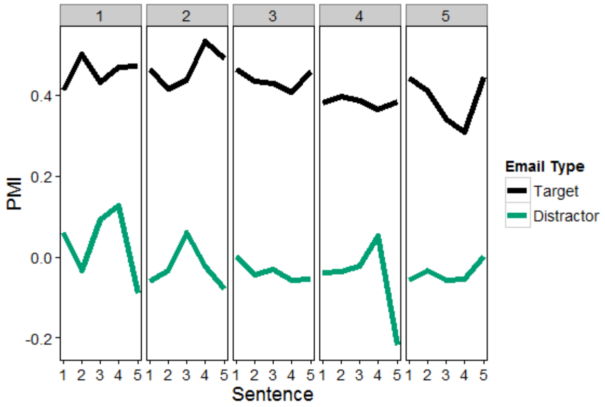

Figure 3. Mean pointwise mutual information (PMI) of each word as a function of position in the document, colored based on the topic.

This decision model has previously been applied to this data set (Harbison et al., 2015). However, the present application differs in that instead of using the random walk process by drawing randomly from target and distractor distributions, we used the skimming strategies specified above to systematically move through the documents and use the evidence collected as the basis for making the decision. That is, instead of a random walk through possible evidence values, we modeled the decision process as a systematic walk through the evidence that resulted from the skimming process. Therefore, the skimming and decision process are guided by the information ecology of the environment.

## Participant Data

The data used here is drawn from a larger experiment on triage decision making that examined response type for two different topics (Harbison et al., 2015). Here we used the results of only one of the two topic conditions, California Analysis, for all response types. The response types include a continuous rating condition (Rating) in which participants responded to an item's relevancy along a sliding scale, and two binary conditions: one instructed to minimize false alarms (FA) and one instructed to maximize the number of targets found (Hit). Further details are below.

The results included 120 trials from twenty-four students in the Rating condition, twenty-six in the Hit condition, and twenty-seven in the FA condition. The trials were a mix of 60 emails about the target topic and 60 emails that were not. Each email was presented one at a time in a random order unique to each participant. Participants were instructed to imagine that they were an investigator looking for information about the Enron scandal, and they were given an excerpt from a newspaper article to read about the scandal (Berenson, 2002). A brief description of the target topic was visible on screen throughout the experiment. Responses were required to respond before the subsequent email was displayed.

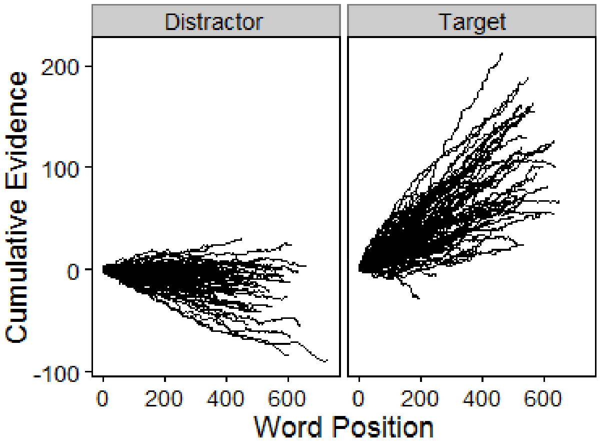

Figure 4. Running sum of the PMI for each of the distractor and target emails.

The email response portion of the study was limited to 70 minutes, with enforced one-minute breaks at 17.5 minutes, 35 minutes, and 52.5 minutes. This gave participants only 35 seconds on average to respond to each email, which forced them to spend some amount of time skimming rather than thoroughly reading every document. Participants were informed of the time constraints, which were designed to force them to make rapid judgments, as in a real-life triage decision making process.

Participants were paid a base rate of $10 in addition to a bonus of $0-15 based on task performance. The calculation of the bonus varied by condition. Prior to the main experiment, participants saw two practice trials, one relevant and one irrelevant. They received feedback on their responses to these trials and were also shown (1) what they would have earned on those trials had they been real, and (2) what other possible responses on those trials would have earned. During the main task, participants were not given feedback on their performance.

Participants in both of the binary conditions made their responses by clicking on either the "Not Relevant" or the "Relevant" button. In both conditions, participants made no additional money for incorrect classifications (i.e., responses of "Relevant" to a distractor email or "Irrelevant" to a target email). However, the amount earned per correct response varied as a function of condition. In the Hit condition, participants were paid $0.14 for correctly identifying target email and $0.11 for correctly rejecting an irrelevant item. Participants in the FA condition received the inverse reward, $0.11 for hits and $0.14 for correct rejections.

In the Rating condition, participants gave ratings using a slider, with endpoints labeled "Completely confident it is not relevant" and "Completely confident it is relevant." The slider response was translated into a rating between 0 and 100, with 0 corresponding to the left endpoint and 100 to the right endpoint. Participants earned between $0 and $0.125 per email rated. Earnings were determined by the spherical scoring formula (Merkle & Steyvers, 2013):

$$earning = \frac{0.125 * r_i}{\sqrt{r_i^2 + (1 - r_i)^2)}},$$

where $r_i = d_i * f_i + (100 - d_i)(100 - f_i)$, and $d_i$ is the relevance of the current trial (100 if relevant, 0 if irrelevant) and $f_i$ is the participant's reported confidence that the trial is relevant.

This formula awards $0.125 for responding "Completely confident it is not relevant" to distractor emails and for responding "Completely confident it is relevant" to target emails. It rewards $0 for the inverse responses, and it awards amounts between $0 and $0.125 for responses in the middle of the scale, with the amount depending on the relevance of the email and the location of the response. Importantly, this formula incentivizes participants to accurately report their confidence that the email is relevant. That is, a participant maximizes their expected reward by responding on the point on the slider that corresponds to their actual subjective belief that the email is relevant (Merkle & Steyvers, 2013).

We are primarily interested in the relationship between the model's predictions and the participant data. However, there is one relationship within the data that is important to note. There was a negative relationship between the magnitude of participant responses and the time to make the response or reaction time (rho = -.288, p<.01). The longer participants took to respond, the closer that response was to the middle of the scale. This relationship between mean response magnitude and mean response time was also found in the two binary conditions (Hit: rho = -.443; FA: rho = -.374, p's<.001).

## Simulation

We tested the model on the same emails that participants read and compared the ratings of relevance supplied by participants with the amount of evidence collected by the model. Table 1 shows three correlations for each skimming strategy. The first is the correlation between the response of the model and the reaction time or time taken for the model to make the response. The second and third correlations are between the responses made by the model and by the participant and the time taken for the model and the time taken by the participant to make a response. Note that the simulation results were not from an exhaustive exploration of the parameter space, but instead from a limited search. The decision threshold for both the skimming strategies was 20 (+20 for responding relevant and -20 for responding irrelevant). The skimming parameters were a stopping threshold ($\delta_S$) of 20, the size of the increment was 3, and size of the change necessary to not be considered an information search failure ($\delta_F$) was .5 for the Skim Beginning Strategy and $\delta_S$ of 1, increment of 1, and $\delta_F$. of 5 for the Skim Paragraph Strategy.

For all three strategies there was a positive correlation between the model's evidence value (or the model's version of a response) and participant responses. The correlation was slightly smaller for the Skim Paragraph Strategy relative to the other strategies. There was also a positive correlation between participant response time and the response times predicted by the three strategies, with a slightly larger correlation between the data and the Skim Beginning Strategy's RTs. Lastly, we looked at the correlation between response time and response magnitude. We found a negative correlation between response magnitude and response time in participant data. Only the Skim Beginning Strategy predicted this negative correlation. The other two strategies incorrectly predicted a positive relationship between response magnitude and response time.

Table 1: Correlations between data and model.

| | | Cor. With Data | |
|---|---|---|---|
| | Resp. and RT | Response | RT |
| Complete | .29* | .58*** | .21* |
| Skim-Beg | **-.46*** ** | **.58*** ** | **.29** ** |
| Skim-Para | .42** | .49*** | .21** |

* p<.05, **p<.01, ***p<.001

We applied the most successful strategy, Skim-Beginning, to the two binary conditions, the results are shown in Table 2. The correlation between the mean participant response per email to the model's predictions were both greater than .50 and the correlation of the reaction times was correlated at greater than .30 (p's < .001).

Overall, the Skim Beginning Strategy most clearly approximates participants' actual responses and response times. Although all three strategies showed the same pattern of correlations with participant responses and reaction time. Only the Skim-Beginning Strategy yielded the negative correlation between response magnitude and reaction time that was found in all three conditions of the data. The model, with this strategy, also correlates with participant responses and reaction times in the two binary response conditions.

## Discussion

Triage decision making is increasingly relevant given the growing amount of information in any given field and the limited time and cognitive resources available to process this information. How are these decisions made? In the present paper, we presented new experimental data on the triage process as well as an initial model. This model combines what we have learned about information search termination decisions (the incremental stopping rule), a skimming strategy (skim from the beginning), a method of representing the information value of words (pointwise mutual information), and a model of decision making and confidence judgment (2DSD) to account for human decisions about the relevance of emails to a target topic.

The general performance of the model matches participant performance. It shows a moderate correlation between predicted and observed ratings of relevancy, as well as between the predicted ratings and the average triage decisions. We also found a correlation between the model's predicted reaction times and participant reaction times. Finally, the model, when using the skim-from-the-beginning skimming strategy, demonstrated the same negative correlation between response magnitude and reaction time as found in the participant data.

Table 2. Correlation between the model and the binary conditions

| | Hit | FA |
|---|---|---|
| Response | .524*** | .555*** |
| Reaction Time | .302*** | .352*** |

*** p <.001

One potential problem with the current representation of information distribution, particularly the relationships shown in Figure 3, is that it ignores the potential of redundancy between words. The information value of each word in each sentence was treated independently, without regard to the words that preceded it. It is doubtful that the evidence provided by each word would be treated completely independently. Just as words are more or less common given a specific topic, words are expected to show a similar relationship to each other. Words are likely related to each other such that processing one would provide evidence about the likelihood of another. This type of redundancy would cause diminishing returns from the continued skimming of a document. This could be accounted with the PMI calculation that included not only the relationship between a given word and the topic, but also the words with each other. We plan to explore this possibility in future work.

## Acknowledgments

This research was supported by the University of Maryland Center for Advanced Study of Language with funding from the Department of Defense.

## References

Berenson, A. (2002, May 9). Mystery of Enron and California's Power Crisis, New York Times.

Berry, M. W., & Browne, M. (2007). 2001 Annotated (by Topic) Enron Email Data Set Documentation. Linguistics Data Consortium. Retrieved from https://catalog.ldc.upenn.edu/docs/LDC2007T22/

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55,* 77-84.

Duggan, G. B., & Payne, S. J. (2009). Text skimming: The process and effectiveness of foraging through text under time pressure. *Journal of Experimental Psychology: Applied*, *15*, 228-242.

Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. *Cognition*, *111,* 397-402.

Harbison, J. I., Mishler, A., & Wallsten, T. (2015). Triage decision making task: An analyst-relevant outcome measure. College Park, MD: CASL.

Hutchinson, J. M., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: can a generalist adapt its rules to dispersal of items across patches? *Animal Behaviour, 75*, 1331-1349.

Merkle, E. C., & Steyvers, M. (2013). Choosing a Strictly Proper Scoring Rule. Decision Analysis, 10(4), 292-304. doi: doi:10.1287/deca.2013.0280.

Pleskac, T. J. & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117,* 864-901.

Xu, G., Yang, S. H., & Li, H. (2009). Named entity mining from click-through data using weakly supervised latent dirichlet allocation in *KDD*, 2009.