The critical moment is coming: Modeling the dynamics of suspense

Zhi-Wei Li (zhiwei.li@nyu.edu)
Center for Neural Science, New York University

Neil R. Bramley (neil.bramley@ed.ac.uk)
Department of Psychology, University of Edinburgh

Todd M. Gureckis (todd.gureckis@nyu.edu)
Department of Psychology, New York University

Abstract

Suspense is an affective state that contributes to our enjoyment of experiences such as movies and sports. Ely, Frankel, and Kamenica (2015) proposed a formal definition of suspense which depends on the variance of subjective future beliefs about an outcome of interest (e.g., winning a game). In order to evaluate this theory, we designed a task based on the card game Blackjack where a variety of suspense dynamics can be experimentally induced. By presenting participants with identical sequences of information (i.e., card draws), but manipulating contextual knowledge (i.e., their understanding of the rules of the game) we were able to show that self-reported suspense follows the predictions of the model. Follow-up model comparison further showed an advantage for the "suspense as variance of future beliefs" account over a number of alternative definitions of suspense, including some that depend only on current uncertainty (not the future). This paper is an initial attempt to link aspects of formal models of information and uncertainty with affective cognitive states.

Keywords: suspense; affect; prediction; expectation; probabilistic modelling

Introduction

Suspense refers to sensations of hopeful or anxious anticipation. These familiar affective states often precede the revelation of important information—exam results, paternity tests, election outcomes and so forth. However, we also feel suspense in situations where there are no direct personal consequences. For example, children enjoy listening to stories that happen in imagined kingdoms, adults spend time watching televised sports, and Hollywood movies are a multi-billion dollar industry. A key feature of these experiences is that information is incrementally revealed over time to the observer, often with the goal of building anticipation and suspense. The goal of this paper is to empirically study the relation between self-reported feelings of suspense and the dynamics of information and uncertainty.

Suspense as the variance in future beliefs

A recent theory in the economics literature proposes that suspense can be explained as an increasing function of the "variance of future beliefs" (Ely et al., 2015). Here the beliefs refer to the probability of a significant outcome (e.g., which team will win a game) that is updated in time with information as an experience unfolds. People are assumed to also estimate how their belief may change in the future. For example, if a doctor arranges to call a patient at a particular time with test results, in the period leading up to the phone call the patient might expect that their belief about their health could soon change (although they may not know what they will learn). Conditioned on the information one expects to receive, if the subsequent future beliefs would be very different from one another they would be said to have high variance. For example, if the test the doctor performed was routine, the patient would not expect their future knowledge state to change much after the call (low variance). As a result they would experience low levels of suspense. In contrast, if the test was a cancer screening, then the call might either alter the person's life or leave them reassured (high variance), and thus they would experience high levels of suspense in that moment.

To formalize these intuitions, we assume belief change is Markovian in that a viewer's subjective belief μ about some outcome evolves over a series of discrete time points t, such as individual points in tennis, card draws in a game, or time passing in a movie. At each time point, relevant information may be encountered and people update their beliefs μ_t (e.g., by Bayesian updating). In addition, viewers also anticipate future information using their understanding of the situation. For example, a viewer might anticipate that their favorite team will score on the next play or that the opposing team will score, each representing a state s. The state s has a probability of being realized $P(s)$ and will result in a future belief μ^s_{t+1}. The variance among these beliefs indicates how different the future might be, and therefore how much suspense might be evoked.

Formally, Ely et al. defined the momentary suspense at time t, S_t as:

$$
\begin{aligned}
S_t &= \mathrm{E}_s[(\mu^s_{t+1} - \langle \mu_{t+1} \rangle_s)^2] \\
&= \mathrm{E}_s[(\mu^s_{t+1} - \mu_t)^2] \\
&= \sum_s P(s)(\mu^s_{t+1} - \mu_t)^2
\end{aligned}
\tag{1}
$$

and we adopt the same notation throughout this paper.

Note that the term $(\mu^s_{t+1} - \mu_t)^2$ may be also interpreted as a metric of variance of belief change or "surprise" that follows learning a piece of information. As a result, the value, S_t can be also be interpreted as the expected future surprise or expected future belief change from the next time period.

Figure 1 gives a graphical overview of the model applied to a hypothetical tennis game. Here μ is the probability of winning the game ($\mu = 1$ if team A wins and $\mu = 0$ if they lose), each point is one time step, and s is whoever wins the

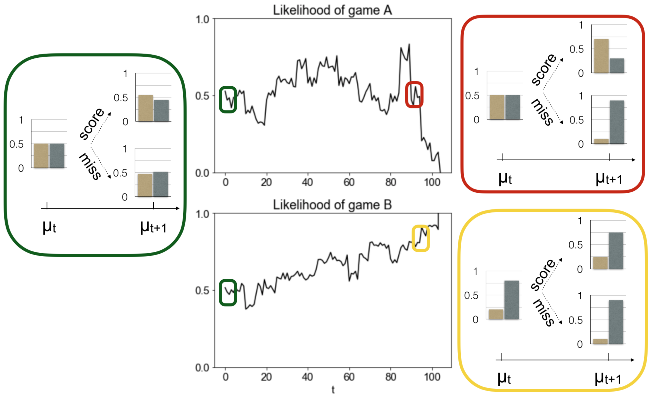

Figure 1: Demonstration of the belief trajectory during watch tennis games and the related suspense predicted from our model. Details see the main text.

next point. In the center of Figure 1 we show the unfolding of belief about who will win for two different games (A and B) with the x-axis representing time. The beginning of both games is not very suspenseful, since whoever wins the first few points has little impact on predictions about the final outcome. However, the end of the game A is more suspenseful since whoever wins a point will greatly swing the final outcome, while game B is less suspenseful since one side has already virtually secured victory.

An experimental test of the theory

Ely et al. (2015) articulated the basic outline of the theory described above and explored a number theoretical analyses of the optimal structure for games to maintain suspense. However, to our knowledge, this operational approach to suspense has not yet been examined empirically. We propose that a useful behavioral paradigm for testing this theory needs to have at least two features:

1. The experiment context should be quantifiable in a probabilistic model. This tends to exclude tasks like reading stories and watching movies because it is not trivial to convert these complex situations into accurate probability models.
2. The experiment paradigm should allow the decoupling of the external stimulus and internal belief. In most prior work, changes in suspense are always confounded with incidental features of the stimuli. To validate the belief-based account of suspense, the ideal experiment would manipulate an observed internal belief through some prior knowledge while holding other aspects of the stimulus and task identical.

With these criteria in mind, we designed a card game related to the classic casino game Blackjack. Participants are asked to draw cards from a small deck with a known distribution of cards and report their moment-by-moment suspense. Intuitively, suspense builds in the task when the sum of the drawn cards approaches a boundary value (21 in Blackjack). If the sum exceeds or hits this value the game is lost. Because the distribution of cards and the probability of drawing any card can be determined exactly, the game is an ideal test bed for exploring information-theoretic models of suspense. In addition, the game is relatively fun, intuitive, and easy to explain to participants.

To address the second concern from above, participants were given one of two different rules for how the game would be scored. In one version, the game was lost anytime the sum of the cards drawn so far met or exceeded the boundary value. This is the traditional concept of "bust" from Blackjack. In a second version, the game was lost only if the sum met or exceeded the boundary value on the final draw of the game. Due to the presence of negatively valued cards, it was possible for the sum to exceed and then return to safety. The differences between these two rules allows us to compare identical sequences of cards, but to modulate if a given card draw was more or less suspenseful about the game outcome according the the Ely et al. theory. To optimize the power of our design, we used a computer-aided method to search for best rules, card decks, and card sequences that result in strong predicted suspense differences under the two rules.

Methods

Participants 263 people (113 female), age 36.7±20.4 (mean±SD) were recruited from Amazon Mechanical Turk using psiTurk (Gureckis et al., 2016) and paid 90 cents (60 cents of this was a bonus that in actuality was the same for all participants). The task took 12±3 minutes to complete.

Procedure Participants were told that we were interested in their feelings of suspense while playing a simple card game. Each participant went through an extensive tutorial covering the rules of the game, and could only continue if they correctly answered a series of comprehension questions. They then played two rounds of training games which were identical to the real games except they were told there would be no bonus. After completing these tasks, participants played a sequence of three games with a $0.60 bonus payment for each game that was won (as describe below, all participants won one game). Afterwards they answered a questionnaire about their strategies and about their perception of the task.

Similar to Blackjack, in each round of a game, the player draws cards from a deck of nine cards. To increase the trial-by-trial suspense dynamics, we use a two-step process for choosing each card: first, the participant sees the animation of nine cards shuffling (Figure 2A); next, the first two cards at the top of the deck were selected (Figure 2B); next, the participant uses the keyboard to spin an animated wheel which (depending where it lands) decides the identity of the final card (Figure 2C). The wheel was programmed so that it spins more when the participant presses the key longer but, unknown to subjects, the spinner always ends up selecting a predetermined card. The purpose of the spinner was to give

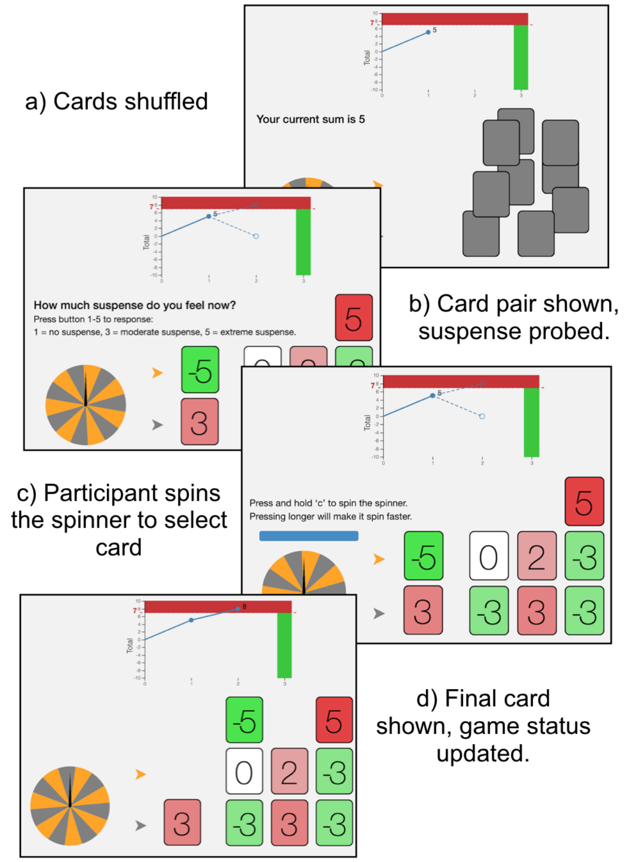

Figure 2: The game interface. a) Card shuffle is an animation that the participant cannot control. b) Participants see the selected card pair and are probed about their suspense on a scale of 1-5. c) Participants press and hold a key to spin the animated wheel. d) When the wheel stops one of the two selected cards is chosen. The whole process repeats until the game has ended.

participants a feeling of control and chance thus they do not lose interest early, although in fact the sequence of cards to be chosen was fixed for the purposes of experimental control. After a card is selected the participant's current card total (the sum of the face value of all of the cards they have drawn so far) was automatically updated in a graph at the top of the screen (Figure 2D).

To measure suspense, after the two candidate cards are shown and before spinning the wheel, we directly asked the participant to rate their current suspense with the keyboard from number 1 to 5 where 1 means no suspense and 5 means very suspenseful. Previous studies on suspense have also chosen a 7-point scale (Gerrig & Bernardo, 1994; Knobloch-Westerwick, David, Eastin, Tamborini, & Greenwood, 2009) and 11-point scale (Comisky & Bryant, 1982; Cupchtk, Oatleyb, & Vorderee, 1998), yet we are unaware of any systematic comparison of different response scales for suspense measurement. No other instructions were given about the use of the scale. However, we asked participants to report how they personally defined suspense in the post-task questionnaire.

Implementing the belief updating model To calculate the belief μ (probability of winning) at a given moment t, we use an exact enumerative strategy. We first enumerate all the possible future card draws remaining in the game according to the known card distribution of the deck. Summing these values, we get the predicted card sum probability. The winning probability calculation is rule-dependent: if the game is played according to the bust rule, we get the card sum distribution for one future step, keep the surviving card sums, continue to the next step and so forth until the game end. If the rule is no-bust (i.e. only the sum of cards at the end of the game matters), we directly calculate the card sum distribution at the end of the game and count the proportion of winning relative to losing sums.

Since the suspense is reported after the pair of possible cards are shown, we assume that suspense is the variance of future probabilities of winning after spinning the wheel and the card being finally drawn. Given that the wheel has equal area for both options, the probability of both future states are equal: $p(s) = 0.5$. The suspense prediction can then be calculated utilizing the equation 1.

Design We will introduce the design of card sequences, then the condition and counterbalance structure.

Belief manipulation: Model-based stimuli design. One key aspect of the theory is that suspense is the result of an active prediction about future stimuli and future beliefs, not the mere reaction to current stimuli. To test this, we looked for rule-dependent differences in suspense responses for the identical card sequences. Given the inherently noisy nature of self reports, we looked for sequences with large predicted differences by maximising a score:

$$\begin{aligned}\text{score}(\text{seq},\text{deck},\text{rulepair}) = S^{\text{rule1}} + S^{\text{rule2}} \\ -\alpha \cdot r(S^{\text{rule1}}, S^{\text{rule2}})\end{aligned} \tag{2}$$

where α is a positive weight constant and $r(\cdot)$ is Pearson's correlation coefficient. The first two terms ensure the average suspense level is not too low while the third encourages anti-correlation between the suspense trajectory under two rules. We set α to a positive constant that makes the two terms have similar magnitude.

We searched the space of rules by generating 5000 random combinations of deck and card sequence valid under both rules and scoring them, then filtered with restrictions to ensure the game also feels like plausible random draws from the deck (details on *Github*). The result of this search was a set of 3 deck/card sequence combinations that evoke strongly different suspense trajectories under two rules: *Bust* with a bound of 7—i.e., the card sum should never exceed 7—and *No-bust* game with a bound of 3—i.e., the sum of cards should not exceed 3 at the end. The full sequences are shown in in Figure 3a.

Participants were randomly assigned to one of the two rule conditions. Two of the games were selected from *high suspense 1-3*.

Besides sequences with interesting suspense dynamics, we also designed two *no suspense* games where the card pairs have similar or identical values, or values that are non-consequential to the games outcome (Figure 3, ”no suspense” 1-2), thus should intuitively induce low suspense. According to Ely et al. model, the predicted suspense at every point in these games is zero.

Task duration and manipulation. Each participant was assigned to one rule condition (rule was a between subject manipulation) and played two rounds of training games (with no bonus regarding the game consequence) then three rounds of gambling games. This is to make the task short enough to avoid boredom. Among the three rounds, two are of *high suspense* and one was a *no suspense* game. The order of games were all counterbalanced. The sign of cards and requisite bound values were also counterbalanced (for example, “cards sum must not exceed 3” was flipped to “card sums no smaller than -3” for half of participants).

Results

Given that each subject may use the scale differently, we *z*-scored the raw suspense ratings for each subject for all analyzes except the likelihood analysis. We also collapsed across the counterbalanced conditions of positive and negative card values. Figure 3 shows a detailed summary of the model predictions and point-by-point empirical suspense ratings for each of the games.

To first assess if the *no suspense* and *high suspense* game type altered people's ratings we ran a paired t-test for each participant's averaged suspense rating from the *no suspense* vs *high suspense* games. The suspense level in the *high suspense* games (0.1 ± 0.3, Mean $\pm$ SD) is significantly higher than the *no suspense* games (-0.7 ± 0.7): $t(262) = 14.18, p < .001$, verifying the basic effectiveness of this very heavy-handed manipulation. Visual inspection of Figure 3 confirms this as well. Participants responded with the lowest increment on the scale for 71.9% and 53.4% of the two *no suspense* games.

To study the direction and magnitude of suspense differences for identical card sequences under different rules, we computed the average z-scored rated response for each point in each of the high suspense games and calculate the difference between in the two rules, comparing this empirical difference to the difference in suspense generated by the model. In Figure 4 we see that most point differences are in the same direction (quadrant 1 and 3). The self-reported suspense difference has an correlation coefficient of $r = 0.80$ ($p = 0.01$) with the model with zero free parameters which is impressive given the inherently noisy measurements of self-reported suspense.

Alternative models

So far we have focused on the formulation of suspense proposed by Ely et al. (2015). In this last section we explore alternatives that may also capture the empirical patterns in suspense.

Alternative probability distance metrics To measure the expected belief change, Ely et al used a squared distance between probabilities while alternative metrics such as information gain and absolute change are common in other contexts (Nelson, 2005). It is unclear in the context of suspense judgment which metrics will best describe people, thus we explore these alternatives.

In the Ely et al model the suspense is defined with an L-2 norm distance for belief update:

$$S_{\text{L2}} = E[(p_{t+1,i} - p_t)^2] \tag{3}$$

where $i = 1, 2$ for each possible card to be drawn and $E[\cdot]$ denotes the average over i.

We explore alternative metrics to quantify the belief update with a KL norm:

$$S_{\text{KL}} = E[KL(p_{t+1,i}, p_t)] \tag{4}$$

an information gain norm:

$$S_{\text{IG}} = E[\text{IG}(p_{t+1,i}, p_t)] \tag{5}$$
$$= E[H(p_{t+1,i}) - H(p_t)], \tag{6}$$

and an absolute error norm

$$S_{\text{L1}} = E[abs(p_{t+1,i} - p_t)] \tag{7}$$

Uncertainty The second theoretical proposal is that people may feel more suspense simply when they have high uncertainty or the estimated chance of winning is close to 1/2. In studies of drama, to keep the story captivating, it has been proposed that “the protagonist and the obstacles he encounters must be fairly evenly matched” (Mabley, 1972). Also in the realm of psychology, uncertainty has been found to sustain attention since people demand the reduction of uncertainty (Berlyne, 1960). By looking at our post-task questionnaire, we also found that around 10% of participants reported they define suspense with uncertainty (although it is unclear whether they use this term in the mathematical sense).

Uncertainty should be the highest when the probability of winning is 0.5 and lowest when it is 0 or 1. To capture this idea, we use the entropy of the belief distribution:

$$S_{\text{uncertainty}} = \text{H}(p_t) \tag{8}$$

Suspense when close to losing The last alternative theory is that people may feel more suspense if the negative outcome is very likely to happen or the estimated chance of winning is close to 0. Previous studies in film narratives (Comisky & Bryant, 1982) and sports viewing (Knobloch-Westerwick et al., 2009) both empirically found that when there is a bigger

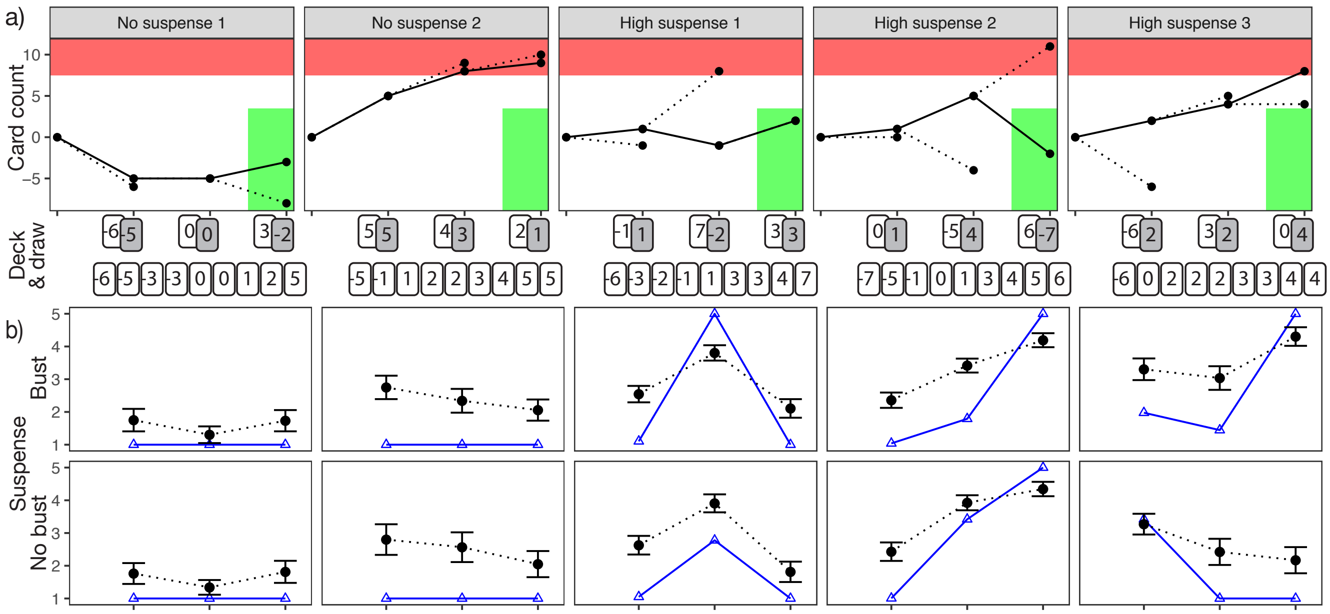

Figure 3: a) Stimuli: Panels show the five games; red indicates the bust region for *Bust* rule and green indicates win region for *No-bust* rule. The card pair at each turn is shown on x-axis with final draw in gray and the full game deck is shown below. Black lines show the actual score and dotted lines show potential score if the alternative card is drawn. b) Results and model predictions: Black circles show M±SE for participants with rule type separated by row. Blue triangles show Ely et al. (2015) predictions scaled to the full response range.

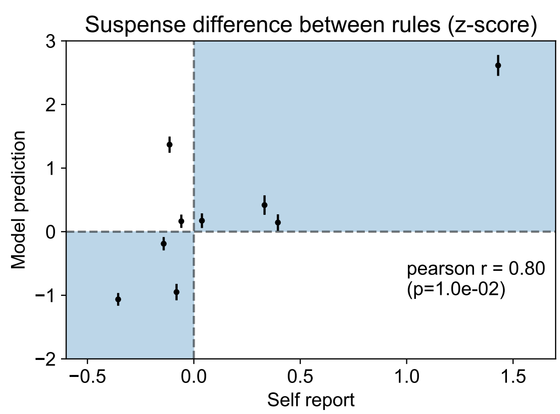

Figure 4: Suspense difference under two rules. Mean±SE differences in z-scored judgments (x-axis) scattered against z-scored model predicted differences (y-axis). Reported suspense differs in the direction the model predicted differences for points in the 1st and 3rd quadrants.

chance for the unwanted outcome to happen, more suspense is felt. In our data we also found this hint: for example, in the two *no Suspense* games, people feel more suspense in *no Suspense* 1 (-0.9 ± 0.5), where the chance of losing is always low than in *no Suspense* 2 (-0.5 ± 0.7) where the card total is always close to the bound and it indeed ends up losing. The difference of average suspense between the two games being significant ($t(261) = -4.98, p < 0.01$) indicates that people may feel more suspense when there is a high chance of losing.

We introduce two models to estimate this “pessimistic” belief about how close one is to losing the game. First, consider a heuristic: how far is the largest of the two cards drawn from the deck is from the boundary:

$$S_{\text{toBound}} = \begin{cases} 1 - |\langle V \rangle_{t+1,i} - \text{bound}|/M \\ 0, \text{if} |\langle V \rangle_{t+1,i} - \text{bound}| > M \end{cases} \quad (9)$$

Where $|\cdot|$ denotes absolute value, $i = 1,2$ representing the card pair and M is the maximum card value (7 in the current design). This piecewise definition assigns zero suspense when the current card sum is too far away from the boundary.

The other model is belief-based which is how big is the probability of losing:

$$S_{\text{pLose}} = \begin{cases} 1 - p_t, \text{if} p_t > 0 \\ 0, \text{if} p_t = 0 \end{cases} \quad (10)$$

$p_t = 0$ represents there is no hope of winning at all thus no suspense.

Likelihood model for fitting discrete responses Fitting the raw suspense scores requires an additional response

Table 1: Model Fits

	Aggregate	Individual	N best fit
L2 (Ely et al)	8.70	0.82	7
L1	**10.06**	**1.00**	**131**
KL	5.97	0.23	13
IG	8.62	0.83	14
toBound	6.65	0.03	40
pLose	8.36	0.20	6
uncertainty	7.47	0.55	52

Note: 1st column: Log likelihood improvement for each fit relative to baseline for average subject judgments (rounded into an integer response). 2nd column: Mean individual log likelihood improvement under optimal shared parameterization. 3rd column: Number of subjects best fit by each model under optimal shared parameterization. Best fitting model indicated in bold

model to convert the continuous suspense predictions to a integer output in the range of 1 to 5. We treat the response as a multinomial sampling process, with the probability of choosing each value related to a beta distribution:

$$p_k = \int_{(k-1)/5}^{k/5} pbeta, k = 1, 2, \ldots, 5 \tag{11}$$

whose beta parameters are defined such that the mean of beta distribution is equal to the suspense prediction (scaled to [0,1]):

$$pbeta(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \tag{12}$$

where

$$a = A * \tilde{S} + 1, b = A * (1 - \tilde{S}) + 1 \tag{13}$$

$A \in [0, \infty)$, and $\tilde{S}$ is the suspense S rescaled to $[0, 1]$.

We define our baseline model where all p_k are equal equivalent to choosing each response randomly. All the model log likelihood results in Table 1 are improvements from this baseline.

For individual participant data we fit this model with A determined by fminbound function of scipy package ($A \in [0, \ldots, 15]$). We compare this maximum log likelihood to that from baseline model and summarize over all subjects. The result of all model comparison is in Table 1.

Our model fitting suggests there is considerable heterogeneity in what drives self-reported suspense in this task. The belief based suspense model with linear belief update distance (L1 norm) fit best overall, suggesting that Ely et al's choice of predictive variance may not be the most natural way of capturing human suspense. However all of the models we considered received some support, with the L2 and information gain models fitting almost identically. Consistent with the self reports in which some participants reported suspense in proportion to their current uncertainty, 20% of individual subjects were best fit by the *uncertainty* model, while a further and 15% best fitted by the heuristic "distance to boundary" model, indicating another potential heuristic sub population distribution.

General Discussion

In this study we designed a paradigm to manipulate the revelation of information about if a player will win a game (and thus earn a monetary bonus) in order to modulate participant's subjective feelings of suspense. We used the model and a computer aided search to select game sequences and rules with high predicted differences in suspense. To our knowledge, this is the first such empirical evaluation of the Ely et al. proposal.

By comparing a range of model variants, we found that most participants were fit by a model that related the rating of suspense to the anticipation of belief change, in line with Ely et al. (2015). However, we also found that belief variability predictions may be better explained by potential absolute (L1) change rather than variance. Heuristic models such as "probability to an unwanted outcome" also captured subsets of the participants.

In sum, this study suggests that suspense is systematically related to meta-cognitive predictions of future belief change. Such preposterior planning (Raiffa, 1974) issues arise in active learning and control contexts. For example, in order to identify the most useful query, one should consider the possible answers one might receive under different possible queries, and how one's beliefs would change as a result (Nelson, 2005). Suspense is thus a quantity that is tantalisingly closely related to such prospective meta-cognition, yet also distinctly low level in that it manifests as a reportable affective state.

Future iterations of our paradigm can be readily adapted to test other interesting hypotheses about suspense, such as the influence of (perceived) control, positive or negative rewards, or the role of suspense in driving attention or engagement (Bezdek et al., 2015).

References

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity.* New York, NY, US: McGraw-Hill Book Company.

Bezdek, M. A., Gerrig, R. J., Wenzel, W. G., Shin, J., Revill, K. P., & Schumacher, E. H. (2015). Neural evidence that suspense narrows attentional focus. *Neuroscience*, *303*, 338–345.

Comisky, P., & Bryant, J. (1982). Factors Involved in Generating Suspense. *Human Communication Research*, *9*(1), 49–58.

Cupchtk, G. C., Oatleyb, K., & Vorderee, P. (1998). Emotional effects of reading excerpts from short stories by James Joyce. , 15.

Ely, J., Frankel, A., & Kamenica, E. (2015). Suspense and surprise. *Journal of Political Economy*, *123*(1), 215–260.

Gerrig, R. J., & Bernardo, A. B. I. (1994, December). Readers as problem-solvers in the experience of suspense. *Poetics*, *22*(6), 459–472.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Knobloch-Westerwick, S., David, P., Eastin, M. S., Tamborini, R., & Greenwood, D. (2009). Sports Spectators' Suspense: Affect and Uncertainty in Sports Entertainment. *Journal of Communication*, *59*(4), 750–767.

Mabley, E. (1972). *Dramatic construction; an outline of basic principles: followed by technical analyses of significant plays by sophocles... and others*. Chilton Book Co.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4).

Raiffa, H. (1974). Applied statistical decision theory.