CHAPTER 2

Expectations about future learning influence moment-to-moment feelings of suspense

2.1 Abstract

Suspense is a cognitive and affective state that is often experienced in the anticipation of information and contributes to the enjoyment and consumption of entertainment such as movies or sports. Ely *et al.* (2015) proposed a formal definition of suspense which relies upon predictions about future belief updates. In order to empirically evaluate this theory, we designed a task based on the casino card game Blackjack where a variety of suspense dynamics can be experimentally induced. Our behavioral data confirmed the explanatory power of this theory.[1] We further compared this formulation with other heuristic models inspired by studies in other domains such as narratives and found that most heuristic models cannot well account for the specific temporal dynamics of suspense across wide range of game variants. We additionally propose a way to test whether experiencing greater levels of suspense motivates more game-playing. In summary, this work is an initial attempt to link formal models of information and uncertainty with affective cognitive states and motivation.

2.2 Introduction

Suspense refers to sensations of hopeful or anxious anticipation. These familiar affective states often precede the revelation of personally 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

[1]The data that support the findings of this study are openly available in the Open Science Framework (OSF) at https://doi.org/10.17605/OSF.IO/KHPR8

multi-billion dollar industry. A key feature of these situations is that information is incrementally revealed over time to the observer, often with the goal of building anticipation and arousal.

Relative to the rich palette of our emotional repertoire, suspense is somewhat unique because it is also associated with a strong motivation or desire for information seeking (e.g., finding out what happens, learning the outcome, etc.). Periods of high suspense are known to modulate arousal and attention mechanisms helping to narrow people's focus to relevant stimuli (Bezdek *et al.*, 2015). For this reason, manipulating suspense is a central concern of the multi-billion dollar entertainment industry. Content producers such as movie script writers, video game designers, and novelists all are trained in techniques to increase and sustain the engagement of consumers by strategically manipulating suspense. However, most of these techniques do not derive from a scientific understanding of the nature of suspense as a human reaction to information.

Recently, Ely, Frankel, and Kamenica (2015) proposed a formal (i.e., mathematical) definition of suspense as being derived from the *expectation* that consequential information will be revealed in an upcoming moment. However, their proposed definition of suspense was entirely theoretical. In the present paper we attempt one very specific goal which is to empirically evaluate the merits of the Ely et al. model as a psychological theory of suspense. We design a novel experimental task that enables us to measure people's moment-by-moment perceptions of suspense. By comparing the predicted suspense from the Ely model (and a number of alternatives) to the responses of participants in our experiment we are able to provide a concrete test of theory.

There are a number of unique contributions of this work. The first is that we empirically test a theoretical model about situations that cause suspense, borrowed from the economics literature, using psychological/behavioral methods. The theory itself is a novel approach for the field of emotion because it allows one to make a-priori predictions about how much suspense a person should feel directly from the context of a task or game. We designed a novel paradigm using algorithmically designed stimuli to quantitatively test the theoretical model. Across two experiments we find support for the general principles of the Ely et al. model, though in a slightly different mathematical form. Inspired by previous research on suspense, we also tested several alternative theories which did not provide as good a fit to the data we collected. We conclude with a third experiment assessing if suspenseful games can drive people to play more games even when they receive less or non monetary compensation, connecting the concept of suspense to information consumption behavior.

2.2.1 Previous research

Suspense is a relatively understudied psychological phenomena. However, there are small, but distinct fields that explore the concept of suspense and the relationship between affective states

and information seeking behaviors. The following section reviews some of this research with the goal of situating the unique aspects of the present work.

2.2.1.1 Theories of suspense

Suspense has been studied in many domains of entertainment including narrative literature, film, gaming, and sports. Confirming everyday intuition, the effect of suspense on enhancing entertainment experience has been empirically verified (narrative: Zillmann 1991; sports: Peterson & Raney 2008; Su-lin *et al.* 1997; gaming: Klimmt *et al.* 2009; advertisement: Alwitt 2002).

However, general principles about what drives suspense remains elusive. In the domain of dramatic story-telling, Comisky and Bryant (1982) propose that suspense will be higher if 1) the audience disposition toward the protagonist is more positive and 2) the belief that the protagonist will fail is higher.

A related concept is the narrative device of removing a possible solution to a problem facing the protagonist (Gerrig & Bernardo, 1994) and situations where a negative event becomes more likely (e.g., in a game, Klimmt *et al.* 2009). Suspense is also known to increase when the audience knows something is about to happen while the character does not (Alwitt, 2002). The temporal dynamics of the narrative also matters. Alwitt et al (2002) propose that the presence of time pressure and alternations between moments of hope and fear before the resolution also increases suspense. The principles identified in these reports are undoubtedly powerful moderators of suspense. However, the deeper principles remain largely qualitative and domain specific.

There have, however, been some attempts to unify the definition of suspense across a broader set of situations or domains. For example, Lehne & Koelsch (2015) attempts to unify the suspense in narrative stories and in music (under the concept "tension"). They developed a domain-independent model stating that suspense "originate(s) from states of conflict, instability, dissonance, or uncertainty that trigger predictive processes directed at future events of emotional significance". The more divergence there is between possible future outcomes, the more suspense is generated. Although a more concrete definition of this divergence is lacking, this theory does highlight some critical psychological components of suspense, particularly the notion of uncertainty and predictive process. en there are diverging possible outcomes for the immediate moment.

2.2.1.2 Information anticipation and uncertainty

While the narrative devices used to build up anticipation in a story are complex and involve many aspects of semantic knowledge, the widely used paradigm of conditioning could be seen as a much simpler form of manipulating an organism's anticipation. Classical conditioning involves the learned anticipation of a positive or negative outcome following an unconditioned stimulus (US),

such as an audio tone which occurs repeatedly before an electric shock (Pavlov, 2010). The period of waiting for the stimulus to arrive may, at least intuitively, involve some of the same emotional feelings of suspense including anxiety and a strong desire to have the uncertainty resolved.

In addition to its simplicity, one advantage of such paradigms is that they allow more careful measurement of how information seeking is modulated by uncertainty and anticipation. Unlike in narrative settings where the uncertainty is hard to quantify, in the conditioning paradigm, researchers are able to manipulate the relationship between cues and rewards thus introducing different levels of uncertainty (White & Monosov, 2016). Monkeys are attracted to visual cues that will resolve uncertainty about future rewards. For example, they are more likely to shift their gaze to informative cues rather than cues signaling more rewards but with no uncertainty (White *et al.*, 2019), indicating the importance of uncertainty reduction for animals. The neural networks responsible for the expected uncertainty resolution have begun to be identified (White *et al.*, 2019; Horan *et al.*, 2019).

Uncertainty-driven arousal and anticipation is not limited to personally experienced outcomes. We experience suspense in movies and stories somewhat vicariously: the fate of the protagonist for instance is not our own. Relatedly, there are a number of experiments exploring behavioral and neural responses to vicariously experienced rewards and punishments (known as social conditioning). For example, if a subject in a conditioning experiment simply observes a video of another subject who they believe is experiencing painful shock, they begin to experience similar levels of anticipatory arousal (Olsson & Phelps, 2007). Similarity and relatedness between the viewer and the foil will enhance the strength of vicarious learning and arousal, e.g., in humans an in-group bias is also present (Golkar *et al.* 2015; see Debiec & Olsson 2017 for a comprehensive review). These findings draw some parallel to the fact that building empathy towards a protagonist is a useful tool for invoking greater suspense.

Besides the similar information delivery structure between conditioning paradigms and suspense-inducing scenarios, the elicited emotion is also closely related. Conditioning is a critical tool for researchers investigating emotions like fear and anxiety in humans (Maren, 2001), that are also closely related with the feeling of suspense (Nomikos *et al.*, 1968). However, as described below, the theory of suspense we develop applies to much more complex situations than the uncertainty about the timing or delivery of a reward or punishment and captures rich temporal dynamics of suspense over time.

2.2.1.3 Non-instrumental information-seeking, or curiosity

Besides being an emotional state, suspense often acts as a motivating force for active information-seeking behavior. Examples already considered include how movies and other media attempt to hold our attention by increasing suspense at key moments.

Conceptually, factors that promote intrinsically motivated information-seeking - sometimes called curiosity (Gottlieb & Oudeyer, 2018; Loewenstein, 1994) also relate to suspense. In line with Berlyne's categorization (Berlyne, 1966), suspense is more linked to "specific exploration" a.k.a information-seeking towards a specific object (as opposed to "diverse exploration", more driven by novelty-seeking, surprisingness, complexity and so on). Suspense is usually about specific questions, like "who will win the game?" or "will the protagonist get killed?" Many recent theories of curiosity reinforce the parallel with suspense. For example, Lowenstein (1994) claims that curiosity is a result of an "information gap." For suspense, the gap naturally exists when the audience cares intensely about the result of something but the information is still not provided yet. Quantitatively, van Lieshout et al. (2018) found that more uncertainty also increases the level of self-reported curiosity as well as eagerness to view an unrevealed outcome.

Despite this work, specific evidence for the impact of suspense on information-seeking is lacking (but see Bezdek *et al.*, 2015). What we do know is that suspense makes sports games (Peterson & Raney, 2008; Su-lin *et al.*, 1997), stories (Zillmann, 1991), and commercial advertisements (Alwitt, 2002) more enjoyable and enjoyment could be a mediating factor for the further information-seeking or consumption. In our study, we will test the behavioral effect of suspense in a more controlled setting, aiming to assess the degree to which experimentally induced feelings of suspense influence the desire to further information consumption.

2.2.2 Theory: Suspense as the expected future belief update

A recent paper proposes that suspense can be defined as an increasing function of the "expected future belief update" (Ely *et al.*, 2015). Here the beliefs refer to the subjective probability of a significant outcome (e.g., which team will win a game) that is updated over time with information as an experience unfolds. For example, while watching a game we might form the impression that there is a 60% chance our favored team will win given the current score and time clock. In addition to tracking their momentary belief, people are assumed to also estimate how their belief may change in the future (a "prospective" type of calculation). 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, following Ely et al., we assume that a viewer's subjective belief μ evolves over a series of discrete time points t, such as individual points in tennis, card draws in a game, or (discretized) time passing in a movie. At each time point, relevant information may be encountered and people update their belief 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 $P(s)$ (determined by things like the mechanics of the game and the abilities of players) 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, $\mathcal{S}_t$ as:

$$\begin{aligned} \mathcal{S}_t &= \mathrm{E}_s[(\mu^s_{t+1} - \mathrm{E}_s[\mu^s_{t+1}])^2] \\ &= \mathrm{E}_s[(\mu^s_{t+1} - \mu_t)^2] \\ &= \sum_s P(s)(\mu^s_{t+1} - \mu_t)^2 \end{aligned} \tag{2.1}$$

The term $\mathrm{E}_s[\cdot]$ represents the expected value of a quantity averaged over all values of s. The step from line 1 to 2 of the equation is because $\mathrm{E}_s[\mu^s_{t+1}] = \mu_t$ (i.e., the belief now is the expectation over all possible future beliefs; derivation in Appendix material).

Note that the difference $\mu^s_{t+1} - \mu_t$ indexes a quantity we might associate with surprise in the sense that it is the difference between what one thinks now compared to after learning a new piece of information.

As a result, the value $\mathcal{S}_t$ can be also be interpreted as the expected future surprise or expected future belief change from the current to the next time period. The phrase "expected surprise determines suspense" is a good summary of the intuitive implications of the theory.

Figure 2.1 gives a graphical overview of the model applied to a hypothetical tennis match. Here μ is the probability of winning the match ($\mu = 1$ if player A is certain to win and $\mu = 0$ if they are certain to lose), each point is one time step, and s is whoever wins the next point. In the center of Figure 2.1 we show the unfolding of belief about who will win for two different matches (1 and 2) with the x-axis representing time. The panel on the left shows why the beginning of both matches is not very suspenseful: whoever wins the first few points has little impact on predictions about the final outcome given how much time remains in the match. However, the end of match 1 is predicted to be more suspenseful since whoever wins a point will greatly swing the final outcome (indicated by the top right panel where μ_{t+1} is quite different, or variable, depending on what happens), while

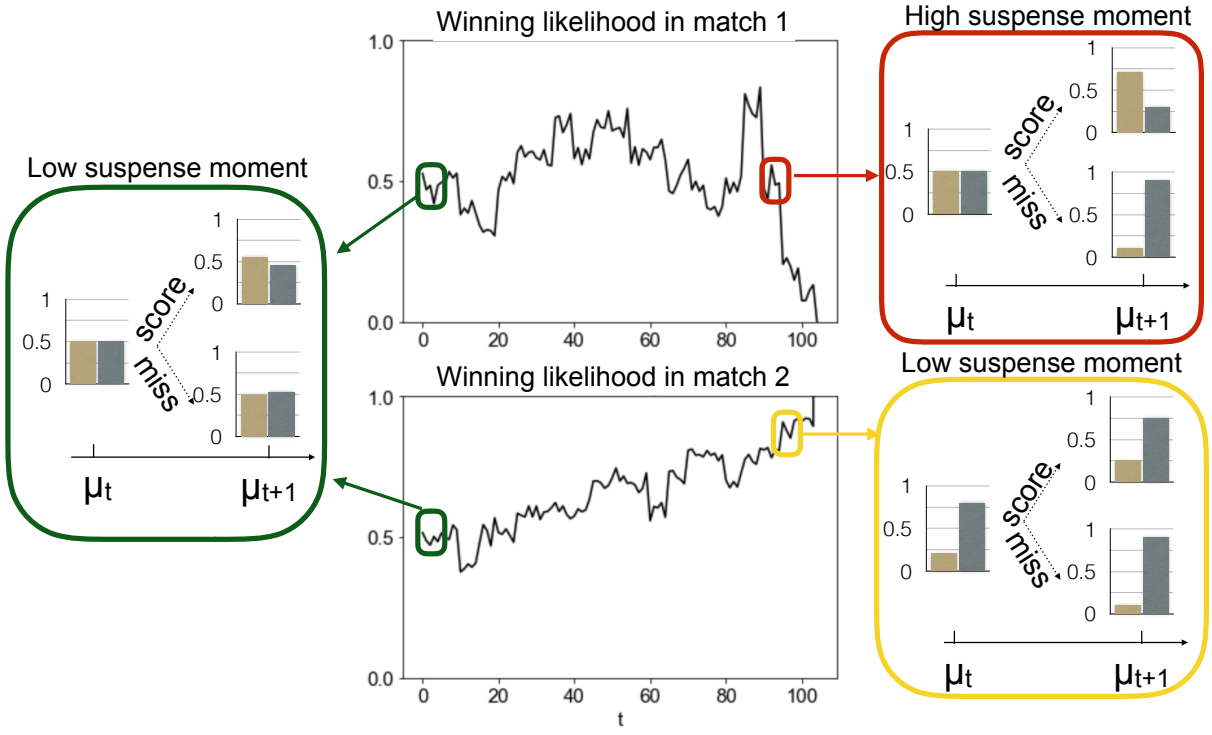

Figure 2.1: The evolution of belief (i.e., likelihood of player A winning) over two different tennis matches. Different moments during the match may correspond to different levels of suspense, that we will explain in terms of a difference in the expected future belief update (colored boxes). **Left (green box)**: At the start, suspense is low because the potential updated beliefs for the next time point do not differ dramatically; **Right top (red box)**: Here is a moment of high predicted suspense in which the next point is expected to have a major impact one way or another on the outcome; **Right bottom (yellow box)**: here player B is already very likely to win, making the expected belief update small thus resulting in low suspense.

match 2 is less suspenseful since one side has already virtually secured victory and so no matter how the next point plays out overall expectations about the match remain the same (bottom right panel, μ_{t+1} is similar no matter what happens). The fact that in match 1 there is more suspense at the end of the match rather than the beginning (despite the overall belief about who will win being somewhat uncertain) shows an important feature of the theory. It is not simply how uncertain you are at the current moment, it is how much you expect the information in the next moment to change your belief.

In the following studies, we will introduce different implementations of this theory, then compare it with other suspense models inspired by the previous research. Details are provided in Experiment 1.

2.3 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 definition of suspense has not yet been examined in psychological experiments (although Wilmot & Keller 2020 conducted an examination using natural language processing techniques). 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. This limitation is more practical than conceptual however.

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. For example, suspenseful moments of a movie might have more visual motion. To validate the belief-based account of suspense, the ideal experiment would manipulate an internal belief through some prior knowledge while holding other aspects of the stimulus and task identical.

With these criteria in mind, we designed a single-player card game related to the classic casino game Blackjack. Participants randomly draw cards from a small deck with a known distribution of cards and report their moment-by-moment suspense. If the sum of cards exceeds or hits a critical threshold the game is lost. Unlike in the Blackjack, the participant does not make decision on when to stop drawing cards but instead is limited by the maximum card number depending on the rules. Thus the participants win by chance. Intuitively, suspense builds in the task when the sum of the drawn cards approaches the critical threshold (e.g., when the sum of drawn card exceeds 21 in Blackjack). 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, including the Ely theory. In addition, the game is relatively fun, intuitive, and easy to explain to participants.

To address the second concern from above, in Experiment 2, participants were given one of two different rules for how the game would be scored, thus matching the stimuli while changing their implications for the outcome. In one version, the game ended in a loss if, at any point, the sum of the cards drawn so far met or exceeded the threshold. This is the traditional concept of "bust" from Blackjack. In a second version, the game ended in a loss only if the sum met or exceeded the boundary value on the final draw of the game. Since we allowed the presence of negatively

valued cards, it was possible, under the second rule set, for the sum to exceed and then return to safety below the threshold before the game ended. The differences between these two rules allow us to compare identical sequences of cards, but to modulate if a given card draw is more or less suspenseful about the game outcome according the the Ely et al. theory. To optimize the power of our experimental approach, we used a computer-aided search to find a combination of rules, decks and card sequences that resulted in strong predicted suspense differences between the two rule sets.

2.3.1 Overview of the experiments

In total, we present 3 experiments using this paradigm. Experiments 1 and 2 test how well the suspense model predicts participants' moment-by-moment subjective report of suspense. Experiment 1 establishes the theory's explanatory power across a relatively large variety of stimuli. At the broadest level, we contrasted games that the model predicts will be extremely low in suspense with games predicted to produce high overall suspense. If the model is even reasonably in line with human suspense, these differences should appear robustly. Second, the model makes point-by-point predictions about the fluctuations in suspense within a game. We used model comparison to assess how well the Ely et al. model accounts for these fluctuations compared against a number of variants heuristics and baselines. Experiment 2 is designed to introduce different temporal dynamics of rules to test the theories in broader contexts. Experiment 3 then asks whether people's willingness to play more games is affected by the level of suspense they experienced in earlier games.

2.4 Experiment 1: Predicting the dynamics of suspense

The goal of Experiment 1 is to provide an initial evaluation of the Ely et al. model. We constructed a wide variety of games and compared the moment-by-moment subjective ratings of suspense from participants with the predicted levels of suspense from the model (and related model variants).

2.4.1 Experimental Method

2.4.1.1 Participants

We recruited 191 participants (age M = 36.2, SD = 13.2, 96 female, 5 undisclosed gender) from Amazon Mechanical Turk using psiTurk (Gureckis *et al.*, 2016). Participants were offered a $0.30 base payment plus the option of a bonus (all participants ended up receiving a bonus between USD

$0.60 and $1.20 as described below). Half of them (96) were randomly assigned to the "high predicted-suspense" condition (described below).

We decided the participant number with the expectation of at least 15 participants per point.

2.4.1.2 Procedure and game design

The instructions and main task were completed by participants on their personal computers using a custom javascript interface in the browser. The task took around 14 minutes (SD=3).

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 to make sure they understood the rules. They then played a training game that was identical to the test games except there was no bonus attached to a win. After completing this, participants played three games. A $0.60 bonus payment was earned for each game that the participant won. As described below, all participants won either one or two games because the outcomes were, in reality, fixed and not under their control, although the task was designed to make it seem to participants that were playing a game of chance.

Similar to Blackjack, in each round of a game, the player drew cards from a deck. In this case, decks contained nine cards with visible values (Figure 2.2A). To increase the trial-by-trial (i.e. point-by-point) suspense dynamics, we used the following two stage process for revealing each card: First, the participant saw the face value of the nine cards in the deck. Then, the cards were flipped over and an animation was shown of the cards being spatially shuffled (Figure 2.2B). Next, two cards at the top of the deck following the shuffle were selected and moved to the left hand side (Figure 2.2C). At this point on a randomly selected subset of trials (around 60%), a self report of suspense was elicited (Figure 2.2D). Next, the participant pressed a button on the keyboard to spin an animated wheel that determined the actual identity of the final card to be drawn (Figure 2.2E). Participants could choose how long to spin the wheel (by holding down a button), but in actuality the spinner always landed on the card determined by our chosen game sequences. The purpose of the spinner was to give participants an (illusory) feeling that the outcome was truly stochastic and that they could potentially use skill of spinner control to obtain a more favorable outcome (increasing engagement with the task). After a card was selected, the participant's current card total (the sum of the face value of all of the cards they had drawn so far) was automatically updated in a graph at the top of the screen (Figure 2.2F). The interface displayed the total sum as well as the graphical history of the sum as it evolved across the sequential draws.

To measure suspense, after the two candidate cards are shown but before the process of spinning the wheel, we asked the participant to rate their current suspense using the keyboard numbers 1 to 5, where 1 means "no suspense" and 5 means "high suspense". Previous studies on subjective

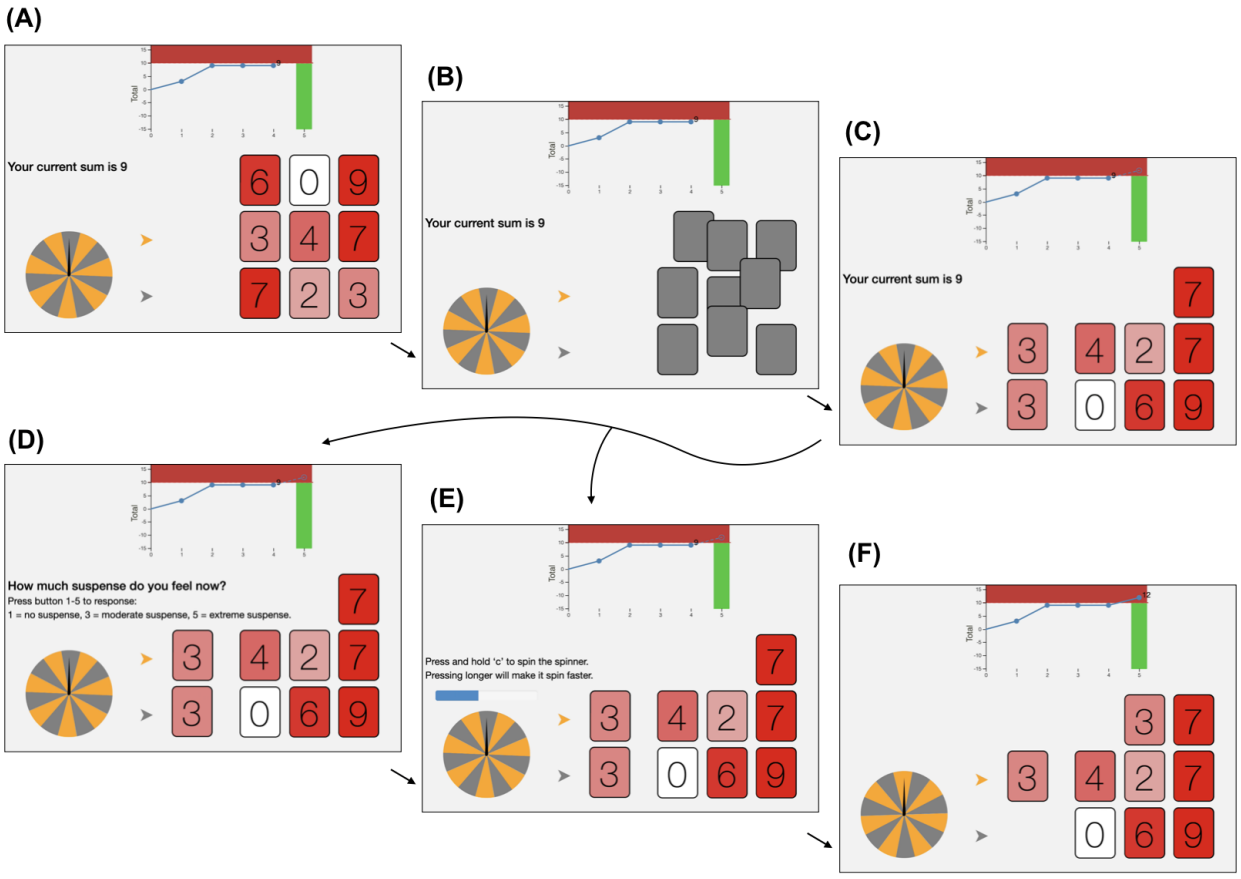

Figure 2.2: The game interface and animation sequence for a single card draw. (A) The deck and current sum are revealed (B) The cards are flipped over and shuffled (with animation) (C) The first 2 cards after shuffling become "focal" candidates (D) Occasionally, self report of suspense are requested from the participant (E) Participants press a button to "control" the spinner speed (F) The final position of the spinner determines which card is finally chosen.

reporting of suspense have also used 7-point (Gerrig & Bernardo, 1994; Knobloch-Westerwick *et al.*, 2009) or 11-point (Cupchtk *et al.*, 1998; Comisky & Bryant, 1982) scales, yet we are unaware of any psychometric comparison of different response scales for suspense measurement. No other instructions were given about the use of the scale. However, we asked participants to describe how they personally understood the term "suspense" in the post-task questionnaire. To minimise interruption of the game experience, suspense rating were only requested on 2 or 3 randomly selected draws per game.

At the end of each game, participants were given the option to play one more game with only half the bonus of the previous games. This was not required and participants were free to decline to play. This final game was intended to test whether the overall amount of suspense encountered in previous games affects people's willingness to play additional games. More comprehensive

explanation and analysis of this data set, along with other contrasting conditions, are presented with Experiment 3.

Finally, participants were presented a questionnaire asking: 1.) If they found practice round helpful for their understanding (range of 1-4), 2.) Whether they felt the game was fair (range of 1-4), 3.) How they judged suspense (free response), 4.) How they decided whether to play the additional game (free response), 5.) If they found any problems or had any suggestions about the whole task (free response), and 6.) Demographic information (gender, age, education). Most of these variables were not analyzed because we had no specific hypothesis (e.g., about gender or age) but they are reported here to facilitate secondary open science data analysis. Participants reported voluntarily and it was still possible to submit the task to Mechanical Turk if the questions were left blank.

2.4.1.3 Stimuli selection

In this task, the potential stimuli space is very large. For a given deck of nine cards, there are 60,466,176 sequences of five draws of two ordered cards with replacement, not to mention practically unbounded variation of deck compositions. This raises the issue of selecting a set of stimuli that will best serve to test the theory.

We searched the space of games by generating 5000 random combinations of deck and card draws that were valid under the game rules, before filtering with several heuristic restrictions to ensure the games also felt like a plausible random draws from the deck. For example, one heuristic restriction prevented drawing the same card more than three times in game. We then calculated the suspense for each draw of each game based on the suspense model (described below) before further filtering the games to find games with either high or low predicted suspense, and both winning and losing outcomes. The full selection procedure is detailed in the Appendix materials.

As mentioned, participants were randomly assigned to either a "model predicted" high or low suspense condition. In the high predicted-suspense condition, all games were selected from the 2% highest suspense games among all the simulated games; in the low predicted-suspense condition, all games were selected from the 10% lowest suspense games. The difference in thresholds for inclusion between the two conditions is because for games end up losing they usually have relatively high level of suspense.

Furthermore, we selected for the outcome of games so that participants experienced either one or two wins out of the three games they played. This ensured that the influence of game outcome on suspense was relatively neutral across participants.

2.4.2 Modeling Approach

2.4.2.1 Belief updating process in the card game

The Ely et al. theory does not explicitly specify the belief updating model that would apply to this game. However, given the simplicity of the game dynamics it is possible to create a Bayesian updating model that is exact. In particular, to calculate the belief μ_t (probability of winning at a given moment), we used an exact enumeration approach, counting all the possible future card draws and their respective win or loss outcomes, which generated an overall win probability. For example, in a game with five total cards draws, if the sum after the fourth card draw is 7 then one can simulate all possible subsequent draws and tally how many result in a win (e.g., card sum $<$10) or loss (e.g., card sum $\geq$10). This determines the probability of going on to win the game given what is known at time point t.

Since the suspense is reported after the pair of potential next cards are drawn, we can calculate the probability of winning (μ_{t+1}) once one of these two cards cards is selected (using the same process as for calculating the current belief). The suspense prediction for this time point then follows equation 2.1. Given the wheel has equal area for both options, the probabilities of both future states are equal (i.e., $p(s) = 0.5$) and so suspense here simplifies to the variance of μ^s_{t+1} over the two possible outcomes s.

2.4.2.2 Model-based data analysis

We introduced a response probability model to convert the continuous suspense predictions from the model (on the range 0.0 to 1.0) to a integer output in the range of 1 to 5. This allowed us to estimate the likelihood of a given subject's response R_t on a given trial t given the predicted suspense $\mathcal{S}_t$ (i.e. $p(R_t|\mathcal{S}_t)$). The response model has a single noise parameter optimized to maximize the likelihood of producing the behavioral data. See the Appendix materials for the full details.

For each individual we collect only 8-9 data points, which is too few for individual modeling, thus we analyzed the data at the group level by using the group parameter for likelihood fitting, and averaging the response given on each point for the correlation analysis. Since each participant was randomly assigned to different games and asked to indicate their suspense at randomly selected points in the game, each averaged data point is based on a slightly different number of participant responses (M = 40.5, SD = 4.0).

2.4.3 Alternative models

Before describing the results of Experiment 1, we consider two types of alternative models: 1) Variations of Ely et al. model that also rely on expected future belief calculation but measure the

probability change in ways other than squared distance (as in Eq 2.1). 2) Heuristic models inspired by previous qualitative research on suspense (introduced in "Previous research" section).

2.4.3.1 Different measures of suspense

To measure suspense as expected belief change, Ely et al. used the squared distance between probabilities before and after encountering some new information. Yet the justification for this choice is somewhat unclear. There are many ways to calculate the how far or how much a belief has changed and so we consider a number of other alternatives (see a Nelson et al. 2005 and 2010 for an extensive discussion of these issues).

To recap, in the Ely et al. model suspense is defined as the expected squared distance, which is also known as a L2-norm, for the belief update on the next time point:

$$S_{\mathrm{L2}} = \mathrm{E}_s[(\mu^s_{t+1} - \mu_t)^2] \tag{2.2}$$

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

We additionally explore alternative metrics to quantify the belief updates. For example, entropy reduction is defined as follows:

$$S_{\mathrm{H}} = \mathrm{E}_s[H(\mu^s_{t+1}) - H(\mu_t)] \tag{2.3}$$

where H denotes the Shannon entropy:

$$H(p) = p\log(p) + (1-p)\log(1-p) \tag{2.4}$$

Alternatively, we could use an absolute error norm, or L1 norm:

$$S_{\mathrm{L1}} = \mathrm{E}_s[\mathrm{abs}(\mu^s_{t+1} - \mu_t)] \tag{2.5}$$

The absolute error norm is closely related to concepts of "probability gain" and "impact" Nelson *et al.* (2010).

A last form is similar to the L-1 norm but a more statistically justified form called the Hellinger distance which is a special case of f-divergence:

$$S_{\mathrm{Hellinger}} = \mathrm{E}_s\left[\mathrm{Hell}(\mu^s_{t+1}, \mu_t)\right] = \mathrm{E}_s\left[\frac{1}{\sqrt{2}}\sqrt{\left(\sqrt{\mu^s_{t+1}} - \sqrt{\mu_t}\right)^2}\right] \tag{2.6}$$

Previous studies provide evidence suggesting human prospective informativeness judgments may be driven by an absolute error norm (Nelson *et al.*, 2010) while there no clear difference between

the other information-theoretic norms (Nelson, 2005).

Intuitive demonstrations of how these mathematical forms differ are provided in Appendix materials (see Appendix Figure 2.13 and 2.14).

Note these different variations are all under the conceptual framework that suspense is based on expected future belief update at the next time point. They are only different in how the magnitude of the belief update is measured.

2.4.3.2 Heuristic models

We now introduce two conceptually different models of suspense inspired by the literature: Uncertainty and "Fear of losing".

Many researchers agree that higher suspense is related to high uncertainty (E.g. Lehne & Koelsch 2015; Although note the "paradox of suspense" Yanal 1996 where uncertainty is not always necessary). Also in the realm of psychology, uncertainty has been found to sustain attention since people may desire the reduction of uncertainty on a motivational level (Berlyne, 1960).

Quantitatively, we define this model in two ways: the entropy of current belief distribution (Uncertainty, probability based) and card difference (cardDiff, pure heuristic), detailed in the appendix.

The last alternative theory we consider is that people may feel more suspense if the negative outcome is very likely to happen. Previous studies in film narratives (Comisky & Bryant, 1982) and sports viewing (Knobloch-Westerwick *et al.*, 2009) both empirically found that when there is a greater chance for the unwanted outcome to happen, more suspense is experienced.

Mathematically, we introduce two models where suspense is higher when probability of winning is lower. These models are identical except that they deal differently with the cases where the probability of losing is certain. Specifically, when probability of winning is zero, it could either be maximum suspense (pLose model) or zero suspense (almostLose model). Thus the almostLose model includes a discontinuity such that the suspense increases as probability of losing increases but falls to zero when losing is certain. The exact formulation and other heuristics related to this idea are detailed in the Appendix "Alternative models" section.

To sum up, we formulated eight models that derive from three different intuitions about suspense as summarized in the Table 2.1. We compare these models to see which best accounts for the behavioral data.

2.4.4 Results

In Experiment 1, we first examined the effect of the high versus low predicted suspense manipulation. This provides a sanity check whether our stimuli lead to clear differences in self-reported

Table 2.1: Alternative Models

Principle	Implementation	Name
Future belief update	Squared belief difference	L2 (Ely)
	Absolute belief difference	L1
	Entropy reduction	ΔH
	Hellinger distance	Hellinger
Uncertainty	Entropy of current belief	uncertainty
	Card difference	cardDiff
Fear of losing	Probability of losing	pLose
	pLose with ceiling	almostLose

suspense. Then, we quantitatively compared the Ely et al. model and other alternatives (all models listed in the Table 2.1) to the behavioral data.

2.4.4.1 High/Low predicted-suspense manipulation

We first checked whether the participants in the low predicted-suspense condition indeed reported lower suspense on average than those in the high predicted-suspense condition. The distributions of suspense responses are clearly separated in two conditions (shown in Figure 2.3). One-sided Mann-Whitney rank test confirms that suspense reported in the high-predicted condition is significantly higher (U = 3.0e5, $p < 0.001$). Participants in the low predicted-suspense condition reported the lowest level of suspense most often, whereas in the high predicted-suspense condition the highest level of suspense was most often selected. Thus, our paradigm had successfully manipulate the feeling of suspense by designing specific decks and card sequences.

Note that in Figure 2.3 the discontinuity in the suspense distribution may due to the instruction about keyboard responding being a little misleading (see Figure 2.2D) which we recommend should be avoided by future researchers).

2.4.4.2 Suspense dynamics

To visualize the detailed game suspense trajectories, we plotted the average suspense reports at each time point in Figure 2.4 (high-predicted suspense games) and Figure 2.5 (low predicted suspense games). To compare these against the predictions of the model variants in Table 2.1, we also plotted the suspense from each class of model: future belief change (L1), uncertainty and fear of losing (almostLose). Note these predictions are parameter-free thus not fit to the data in any way.

From these trajectories we can observe several features. First, there is considerable agreement between participants' suspense judgement such that the aggregate suspense trajectories are not flat

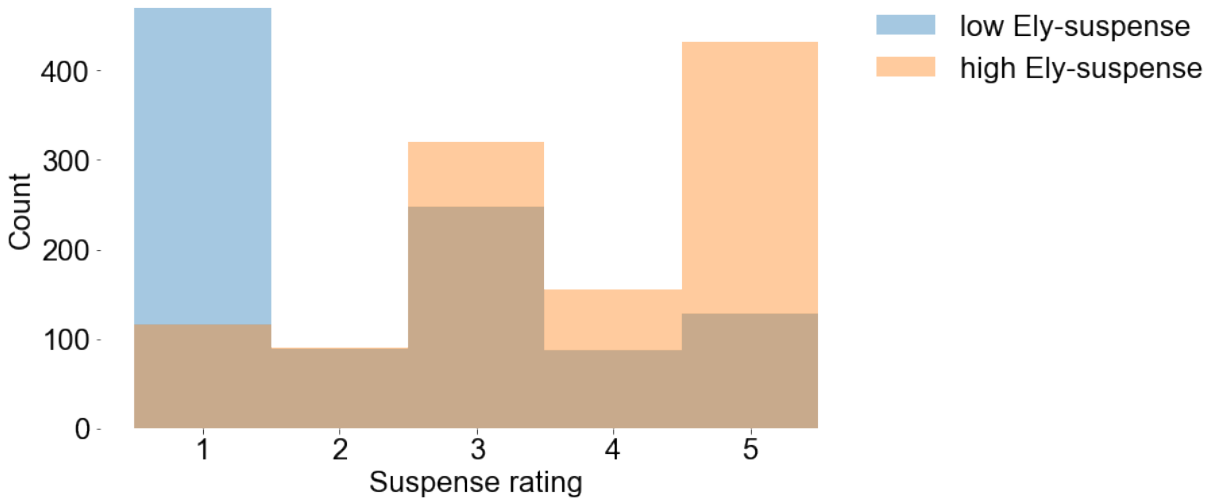

Figure 2.3: Experiment 1, Distribution of reported suspense in low and high predicted-suspense conditions.

Table 2.2: Experiment 1: Model Comparison.

	Future belief update				Uncertainty		Fear of losing	
	L1	Hellinger	L2	ΔH	uncertainty	cardDiff	almostLose	pLose
Pearson's r	0.82	0.82	0.74	0.74	0.74	0.68	0.65	0.54
Likelihood difference	7.47	6.43	3.97	3.85	8.88	3.55	4.56	2.71

Note: 1st row: Correlation coefficient between the parameter-free model prediction and the raw participant report of suspense.
2nd row: Maximum Likelihood Estimate fitting of all individual trials above the baseline (random report), averaged to per trial, in the unit of 10^{-2}.

but shows clear within- as well as across-game dynamics. Second, the L1 model shows qualitatively good tracking of the trends in the high predicted-suspense condition (Figure 2.4). However, in the low predicted-suspense condition, the L1 and uncertainty models systematically underestimate participants' reported suspense level while almostLose heuristic is significantly modulated by if the game leads to a win or loss (Figure 2.5).

To quantitatively test how well the models explain the dynamics of suspense on a point-by-point level, we first checked point-by-point correlation between averaged participant suspense judgments and the parameter-free model predictions across all the games. The correlation coefficients for all models are listed in Table 2.2. The models with the highest correlation are the two "future belief update" models Hellinger and L1 heuristic model, followed by other models in the same family (L2 and ΔH) and the Uncertainty heuristic. Interestingly, all models from the "future belief update" family give higher Pearson r than heuristic models. Figure 2.6 shows the joint distribution

of reported suspense ratings per trial (averaged across people) for examples from each of the three classes of model types.

We also used maximum likelihood estimation to fit a probabilistic version of each model using a single "response noise" parameter to all responses points (N = 2139). Details for how we mapped the model predictions to response scale elements are given in Appendix section ''Computing model likelihood"). The per-trial likelihood improvement for the given model relative a null model which assumes a uniform probability of each response category is presented in Table 2.2 (bottom row). According to this measure, the uncertainty heuristic fits the best, followed by the L1 and Hellinger model. The "fear of losing" models perform relatively worse in both correlation and likelihood. The likelihood fits differs from correlation in requiring a match not just in terms of covariance but also in terms of absolute values of the ratings being calibrated correctly. Also the likelihood fits all the individual trials whereas Pearson'r is against the group average.

2.4.5 Discussion

Our new experimental paradigm allowed us to elicit reasonably consistent suspense reports between participants. The overall model-based high/low suspense manipulation had a clear effect despite the inherent noise in self-reported data.

We then tested several classes of suspense models with specific variants of each. Overall, we found that the "future belief update" models, especially those using the L1 and Hellinger metrics, described the data best. This supports the broad idea at the heart of the Ely et al. proposal that suspense indexes anticipation of future belief changes. However our results are suggestive that the specific formulation of this anticipation (using the L2 norm) may not best characterise human suspense.

Among the (non-expectation-based) heuristics, the current uncertainty model provided the best likelihood fit. It is surprising that the simple Uncertainty model fits the data well since it is insensitive to the whether and when the uncertainty can be resolved. For example, this model assumes that the very beginning of a game can induce th same level of suspense as much as the match point, as long as people are similarly uncertain about the outcome. This seems counter-intuitive and we wonder whether this will generalize beyond the current result since experiment 1 only explored a limited number of games and, more importantly, under a specific game rule.

A key prediction of the Ely et al. approach is that suspense is a function of expected belief rather than something evoked by any particular stimulus or situation. One way to test this hypothesis is to show people games which are identical (in terms of the face value of the cards being drawn) but to manipulate, across conditions, their belief about whether they are close to winning or losing. To this end, in Experiment 2, we designed two games for each of the same deck and card sequence

such that the difference is only in the "rules" of the game that determine when the player wins. This manipulation enables studying the mechanism driving suspense without the potential confound of specific card and deck information. In addition, this provides a replication and a new test of the model's generalizability.

2.5 Experiment 2: Rules with different temporal structures

To disentangle the alternative theories of suspense and examine their performance in a wider variety of game structures, we introduced a pair of alternative game rules:

1. The "Bust" rule. This is same as in Experiment 1 where the game is lost any time the sum of the cards drawn so far meets or exceeds the boundary value.

2. The "No Bust" rule. Here the game is lost only if the sum meets or exceeds the boundary value *on the final draw of the game*. Due to the presence of negatively valued cards, it is possible in this game for the sum to exceed the bound but then return to safety by the final draw.

Using the same card game paradigm except adjusting the rules, we could then contrast the suspense response when playing games with exact same card sequences but under different rules, highlighting how suspense is modulated by the internal belief.

2.5.1 Method

2.5.1.1 Participants

We recruited 263 participants (age M = 36.7, SD = 20.4, 113 female) from Amazon Mechanical Turk using psiTurk (Gureckis *et al.*, 2016). They were paid \$0.90 (\$.30 base pay and a \$.60 bonus which was, in fact, the same for all participants). The task took 12$\pm$3 minutes to complete. There were 144 participants assigned to the "Bust" rule and the rest to "No Bust" rule condition. We decided the participant number with the expectation of at least 80 participants per point. This number is bigger than experiment 1 because we expect the effect of rule manipulation will be much smaller than the high / low suspense game manipulation.

2.5.1.2 Procedure and stimuli

The games and interface were as in Experiment 1 except that there were just 3 draws per game rather than 5. Each participant was assigned to one of the two rule conditions and played 2 practice games then 3 actual gambling games.

2.5.1.3 Rule design and model-based stimulus selection

To optimise the power of the design to distinguish between candidate models, we used computer-aided search to find game sequences (i.e., card decks and the sequence of card draws) which are valid under both rules but also lead to robust differences in predict suspense according to the Ely et al. type belief models.

To implement this design, we searched for a diverse set of games with large predicted suspense differences by maximising:

$$\begin{aligned} \text{score}(\text{seq}, \text{deck}, \text{rulepair}) = \mathcal{S}^{\text{rule1}} + \mathcal{S}^{\text{rule2}} \\ -\alpha \cdot r(\mathcal{S}^{\text{rule1}}, \mathcal{S}^{\text{rule2}}) \end{aligned} \tag{2.7}$$

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 given the Ely et al. model. We recomment setting α to a positive constant around 1 which ensures the two terms have similar magnitude.

We then searched the space of games to select valid stimuli as we did in Experiment 1. For this rule search, we added some additional steps. We searched the rule space for pairs of bust and no-bust rules with different boundary values, then generated random game combinations. We then selected the top 10 games of each rule pair, calculated the average score of these games then picked the rule pair with the best scores. The final winning rules were: 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 of three draws).

Each participant was assigned to one rule condition and played two training games (meaning they did not contribute to the participants' final bonus) then three test games with a potential \$0.30 bonus. Of these, two were overall high predicted-suspense and one was low predicted-suspense. The order of games were all counterbalanced.

2.5.2 Results

The goal of Experiment 2 was to isolate prospective beliefs from other features of our stimuli by matching game sequences to be identical but changing their implications with our rule manipulation. This allows us to better study what drives feelings of suspense even if the game sequences were the same, and how well the different model categories capture that effect. Specifically, we wanted to know if the Uncertainty heuristic model which is not sensitive to the rule's temporal constraints will describe the behavioral data as well.

We first checked the correlation between the game suspense trajectories (averaged over all

Table 2.3: Experiment 2: Model Comparison

	Future belief update				Uncertainty		Fear of losing	
	L1	Hellinger	ΔH	L2	uncertainty	cardDiff	almostLose	pLose
Pearson's r	0.91	0.90	0.86	0.86	0.83	0.77	0.61	0.26
Likelihood difference	11.07	10.64	8.90	8.96	5.96	6.32	2.28	-0.00

Note: 1st row: Correlation coefficient between the parameter-free model prediction and the raw participant report of suspense. 2nd row: Likelihood fitting all individual trials minus the baseline (random report), averaged to per trial, in the unit of 10^{-2}. All the heuristics are worse than the "future belief update" models. The best model is the "future belief update' models of L1 and Hellinger form.

participant response) and the parameter-free model predictions. It turns out that all "future belief update" models have higher correlation as well as likelihood than the other heuristics models (both listed in Table 2.3; Correlation see Figure 2.7). Among the "future belief update" models, metrics forms of L1 and Hellinger again perform the best as in Experiment 1.

In addition, we assessed how well the models predict the suspense difference between the two rule conditions for identical card sequences. This analysis compares the difference in the model prediction for each game point under the two rules against the difference in empirical suspense ratings, with the empirical suspense aggregated among all the participants in the same rule condition. Note the low predicted-suspense games are excluded in this analysis since the suspense in different rules are supposed to be the same.

Figure 2.8 shows the suspense under 2 rules in each game plus their difference. From Figure 2.8C, we can see that although the card sequences and decks are identical, the suspense reported by participants clearly depends on the rule. In several places the difference significantly departs from zero i.e. the 95% confidence interval does not include 0. In those cases, the L1 and Uncertainty models predict this rule difference in the right direction, while almostLose model does not.

Figure 2.9 visualises the match between model predictions and judgments. The L1 and Uncertainty models predict the direction of rule-difference for most game points when the empirical data is unambiguously positive or negative (i.e. the 95% interval does not include 0), as well as the magnitude of rule difference. The Uncertainty model is best able to predict the suspense difference, followed by L1 model, while almostLose model does not work as well. Note the model predictions are generated without any parameter fitting in both Figure 2.8 and Figure 2.9.

2.5.3 Discussion

In Experiment 2 we introduced a novel manipulation approach allowing us to fix the sequences presented to participants but vary their suspense implications by way of two different rules: "Bust" and "No Bust". These two rules were selected to induce different suspense trajectories even though the sequence of card draws was identical. Participants were sensitive to this manipulation, reporting different levels and patterns of of suspense depending on the condition, but the magnitude of this difference is not strong as the model predicted. This could partially be due to the suspense difference are calculated across participants (we did not allow same participant to play through identical games under two rules).

We also found that the future-belief-update model class and uncertainty-based model class more accurately captured the behavioral difference under the 2 rule conditions than the other heuristics.

Comparing to other heuristics inspired by previous studies of suspense, the "fear-of-losing" models were unable to account for the data well, but the current uncertainty-based models performed relatively well, suggesting that both current and prospective uncertainty play a role into suspense judgments. We will give an overall analysis combining all the empirical data in the final Discussion section.

2.6 Experiment 3: Manipulating willingness to play more games

After examining the factors predicting people's perceived suspense, we would like to ask a different question: what are the *effects* of suspense? This is well motivated by Ely *et al.* (2015) where suspense is hypothesized as a non-instrumental utility which people try to optimize upon.

In the context of our card game paradigm, we specifically asked: Are people who have played more suspenseful games also more willing to play more games? Previous studies have suggested that suspense levels correlates with engagement in role-playing computer games (Klimmt *et al.*, 2009), indicating a positive relation between suspense and willingness to play. Here, in a much simpler game will we see a similar effect?

2.6.1 Method

2.6.1.1 Participants

We ran 242 participants (age M = 36.3, SD = 21.5, female 123, 2 undisclosed gender). In the following analysis we will also include the 191 participants from the Experiment 1. For the 242 participants they received no bonus instruction while the 191 participants from Experiment 1 re-

ceived half bonus if they win the additional game. The participant number was expected to be around 200 due the same calculation as experiment 1 and we ended up recruiting more due to a technical error.

2.6.1.2 Procedure

Experiment 3's procedure was the same as Experiment 1 except that, after finishing all games, participants were shown the following message: "All the required games are done, thank you! However, you can also play one more game (with [half/no] bonus)". Then, participants could choose either to stop or to play one more game. If they chose to continue, a random game that they had not experienced before would be presented. We designed the two payment conditions because the monetary reward itself is a critical confounding factor for the behavior being studied, i.e. incentivising continuation orthogonally to the potential intrinsic reward of suspenseful engagement. Exploring two levels of payment helped to avoid the minor modulation of suspense condition being overshadowed by the effect of the monetary incentive.

Notice that a decision of playing more games clearly involves complex considerations: the economic return, the opportunity cost of spending the time playing, as well as any fun derived from the game. To disentangle these motivations, after the full task finishes we asked the participants what was the reason for their decision.

Our hypothesis was that participants assigned to high-suspense condition would be more likely to play one more game than those in the low-suspense condition.

2.6.2 Results

For the 191 participants in the half-bonus variant, we found a slightly higher proportion of those in the high suspense condition chose to play one more game (77 out of 96, 80.2%) than those in the low predicted-suspense variant (70 out of 95, 73.7%). Likewise, in the no-bonus condition, participants in the high suspense condition were more likely to play another game (27 out of 124 (21.7%) compared to 20 out of 118 (16.9%).

To get a Bayesian interpretation of these results, we treated the choice as a random variable drawn from the Bernoulli distribution with a bias rate r then calculated the posterior of bias rate $P(r|data)$ based on a uniform beta prior $B(1,1)$. In the half-bonus variant, the two distributions are separated only by a 55% density interval (Figure 2.10). This is not a statistically significant result (p=0.66 in Fisher exact test). In the no-bonus variant, two distributions are separated by a density interval of 49%. The trend is also not significant in the no-bonus condition (p=0.34 in Fisher exact test).

To combine the two conditions, we performed a logistic regression on the choices with bonus

variant and suspense condition as independent variables, using the statsmodels package in python (Seabold & Perktold, 2010). The coefficient for suspense condition was 0.34 (z=1.42, p=0.16, 95% confidence interval [-0.13, 0.81]), again indicating the higher suspense increases the possibility of choosing to play one more game, though not significant statistically.

To address the concern that the suspense manipulation is not directly reflecting participants' perceived suspense, we also grouped participants by their average self-reported suspense (Figure 2.10, splitting the whole sample to two equal sized groups based on each participant's average reported suspense throughout the games. The result is again non-significant (Half bonus: 51% density interval for distribution separation; no bonus: 52%) but directionally consistent with our hypothesis.

2.6.3 Discussion

Experiment 3 provides some preliminary evidence that suspenseful experiences in a task may enhance people's willingness to engage for longer. This is is in agreement with Ely et al's conjecture that suspense function as a and some previous work in computer games (Klimmt *et al.*, 2009) where they framed a game in qualitatively different ways (with or without threat). Our manipulation of card sequence difference is much more subtle and qualitative, yet still induced certain effects.

This is a surprising effect especially given that we conducted the study online where the task platform provides very little hindrance to task switching thus relatively high opportunity cost for playing additional games with reduced monetary reward. And more importantly, workers on Amazon Mechanical Turk are most strongly motivated by payment, not for fun (Kaufmann *et al.*, 2011). Yet still, in the self reported reason of continuing play games, we saw the fun experience still accounts for participants' choice. We coded the existence of key words to represent two categories of motivation: "win/money/bonus" signaling the drive of extrinsic outcomes, "fun/enjoy" signaling the intrinsic motivation. In the no bonus variant, the majority of participants that continued reported that they did so "for fun" (21,61.8%), and fewer reported they played "to win" (13,38.2%). In the half bonus condition, although most people (81,63.3%) reported their reason as related to winning or getting more bonus or money, there was still frequent mentions of "fun/enjoy" (47,36.7%). This further hints that the fun of game will causally make people play more, and suspense could be one component of that motivation.

2.7 General Discussion

In this paper, we introduced a new paradigm for measuring suspense dynamics. Our task involved a card game designed to lead to diverse situations in terms of current and expected uncertainty. Un-

like previous studies of suspense that have relied on qualitative theories on suspense and relatively coarse manipulations[2], this new paradigm facilitates testing theories of suspense in a quantitative way. Specifically, we tested the Ely et al. (2015) proposal under which suspense is driven by expected future belief change. This theory successfully captured the behavioral data across a range of card sequences (Experiment 1) and captured differences between two game rules for the same card sequence (Experiment 2). Overall, the class of forward looking models shows better generalizability than the alternatives we considered.

2.7.1 Evaluation of models across all experiments

We now combine all the empirical data we have from Experiments 1, 2 and 3 to evaluate the models. Results are summarized in Table 2.4. We fitted the same group noise parameter across all data for each model. Detailed description of the data and comparison are described in the appendix.

The overall winner is the "future-belief-update" using the L1 norm. In terms of correlation with the response averaged across participants, all the "future-belief-update" models have higher correlation coefficients than the other heuristics. In terms of likelihood fitting of each individual's response, the two "future-belief-update" models in L1 and Hellinger forms fit the best, followed by Uncertainty model. The heuristic of "fear-of-losing" performs worst overall. Note that the "unceratinty" heuristics may perform similarly to the "future-belief-update" models in some experimental conditions but not others. Conceptually, this is understandable because the "future-belief-update" models already incorporates part of the "uncertainty" model: these models make more and more similar prediction as the time horizon approaches the end of the task. It is only at the beginning of the game that "future-belief-update" models will predict lower suspense than "uncertainty" model." It is possible that better resolution of these theories might come from larger experiments with a greater variety of stimuli.

Table 2.4: Model comparison for all Experiments.

	Future belief update				Uncertainty		Fear of losing	
	L1	Hellinger	L2	ΔH	uncertainty	cardDiff	almostLose	pLose
Pearson's r	0.82	0.81	0.74	0.74	0.69	0.60	0.65	0.50
Likelihood difference	8.78	7.89	5.92	5.77	6.20	3.85	3.22	0.9

Note: Columns and rows as in Tables 2.2 and 2.3.

Across our three experiments, we thus have good evidence that an anticipation based "future-belief-update" model captures something key to human suspense patterns that cannot be captured

[2]While Comisky & Bryant (1982) do take a somewhat quantitative approach, their manipulation was not entirely numerical. Instead, manipulated text was used qualitative language — "the chance was...": "...absolutely nil" / "...extremely slim at best" / "...somewhat against" / "...roughly even" / "...totally certain"

by present focused uncertainty based models or simpler heuristics. Yet still, we had the concern that people may use a combination of heuristics instead.

Thus, we finally tested a hybrid model allowing the linear combination of the heuristic models of "uncertainty", "cardDiff" and "almostLose". With linear regression against the aggregate response, the hybrid heuristic model gives slightly higher Pearson's r (0.83) than the L1 model (0.82). For the likelihood fitting, however, the hybrid heuristic model gives likelihood 7.68($*10^{-}2$) per trial, which is smaller than the single model of L1 ($8.78*10^{-}2$ per trial) and Hellinger ($7.89*10^{-}2$). Given that the "future-belief-update" model generates suspense prediction without any parameter tuning, this seems to be an empirically promising account of suspense.

2.7.2 Limitations and future directions

This line of experiments has some limitations that will necessitate future investigation. First, although the best model we found was in the family of future belief change models, there is no clear explanation why L1 belief distance metric is consistently preferred. By comparing the different metrics in Appendix Fig 2.14 and 2.13, we found that all the other metrics underestimates the suspense compared to the L1 norm. Whether this preference for belief update judgement relates to other belief-related psychological phenomenon is worth exploring (e.g., does the L1 measure predict judgments of surprise?).

Second, it is likely that participants' understanding of suspense was somewhat heterogeneous and that their subjective access to this quantity was limited. That is, self reports may have been driven by arousal responses from a mixture of sources, including anxiety elicited by uncertainty and fear elicited by potential failure as well as expectations about future belief change. The belief change model class could capture some of these sources while heuristic models may have captured other valenced sources such as fear of losing. Further studies that manipulate outcome valance (either with wins being rare or common, and likely or unlikely given a game state) may help to disentangle these differences.

Third, given the subjectivity of suspense, we used self report. However, self report constantly interrupts the flow of experience. We hope this work could serve as a foundation for further experimental work utilising a more implicit measurement of suspense, for example, EEG signal of suspense.

Fourth, while we took care to construct situations with a wide range of suspense levels, we were still restricted to the space of single-player card games where many potentially important factors of suspense are absent, presumably limiting the ceiling levels of suspense participants experienced. Future work could try to study settings that are known to be particularly captivating yet still simple enough to enable computational analysis. Alternatively, more sophisticated computational tools

can be employed to process more complex information (for example, Wilmot & Keller 2020 used natural language processing techniques to estimate suspense in short stories).

Last, although in Experiment 3 we found evidence suggestive that suspense might increase people's willingness to further engage in a task, the effect was very small and not statistically significant even in our substantial sample size. Besides increasing the range of suspense, we may be able to find more sensitive measurements of the effect of suspense, such as modulation of attention (see Bezdek *et al.* 2015) or willingness to pay to see an outcome. These steps will help provide a normative account of why expected future belief change produces the physiological state that we call suspense.

2.8 Appendix

2.8.1 Stimuli design

In experiment 1, we first simulated 5000 games, then filtered out games that looked "fake", as was complained in pilot studies. Specifically, we excluded games that 1) have certain cards appear $>$ 2 times or 2) have a card pair being selected > 1 times. For the rest of "valid" games, we aimed to present stimuli with different level of suspense. Pilot study suggests that both the Ely et al. suspense and the "pLose" model predict empirical suspense well, we picked a "hybrid suspense" that averaged over these two model predictions as our index. Specifically, we selected the lowest 10% and highest 2% suspense games into our final round of game candidates, including winning and losing games. This asymmetry comes from the fact that for losing games, it was hard to elicit low suspense, thus more games are included for the "low predicted-suspense" group. In our round of simulation, there are 59 low-win, 13 low-lose, 46 high-win, 43 high-lose games. Finally, we randomly selected the number of stimuli we needed for the experiment.

2.8.2 Data from three experiments

Here we delineate the differences of three experiments' suspense report.

The game rules are same in Experiment 1 and 3, but different from Experiment 2. In Experiment 1 and 3, participants each played 3 gambling games, reporting suspense at at most three points, thus contributing 9 data point per person. In Experiment 2, each participant played 5 gambling games, each reporting 3 points.

For likelihood fitting of individual responses, we have 2139 data points from experiment 1, 1879 from experiment 2 and 2367 from experiment 3. For the regression of suspense report across participants, we have 88 game points from experiment 1 (on average 40 participants per point), 30 from experiment 2 (on average 78 participants per point) and 154 from experiment 3 (on average 10 participants per point). Therefore, the cross-participant reports may have unfairly weighted the data from experiment 3 which has more points but also significantly more noise.

2.8.3 Results from Experiment 3

The comparison of models and data are listed in table 2.5.

2.8.4 Computing model likelihood

Given a participant suspense rating R_t on a given trial t (where $R_t \in \{1, 2, 3, 4, 5\}$), we would like to obtain the likelihood of R_t given the predicted suspense of the model $\mathcal{S}_t$ (i.e. $p(R_t|\mathcal{S}_t)$).

Table 2.5: Experiment 3: Model Comparison

	Future belief update				Uncertainty		Fear of losing	
	L1	Hellinger	L2	ΔH	uncertainty	cardDiff	almostLose	pLose
Pearson's r	0.81	0.80	0.73	0.73	0.65	0.52	0.68	0.59
Likelihood difference	7.60	6.43	5.06	4.81	3.85	1.69	3.08	2.84

Note: 1st row: Correlation coefficient between the parameter-free model prediction and the raw participant report of suspense. 2nd row: Likelihood fitting all individual trials minus the baseline (random report), averaged to per trial, in the unit of 10^{-2}. All the heuristics are worse than the "future belief update" models. The best model is the "future belief update' models of L1 and Hellinger form.

We treated the response as a multinomial distribution parameterized by $\mathcal{S}_t$. To determine the exact value of this multinomial distribution, we constructed a beta distribution with its mode being the same value as the model suspense $\mathcal{S}_t$ for each game point. Then, we calculated the cumulative probability density under each 1/5 percentile of this beta distribution, mapping into the integer output of suspense, thus obtaining the likelihood function. An intuitive example of this process is shown in Figure 2.11.

Mathematically, the multinomial can be defined as follows:

$$p_k = \int_{(k-1)/5}^{k/5} \mathrm{Beta}(x;\alpha,\beta)\mathrm{d}x, \mathrm{for} k = 1, 2, \ldots, 5 \tag{2.8}$$

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

$$\mathrm{Beta}(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{2.9}$$

where

$$\alpha = A * S + 1, \beta = A * (1 - S) + 1 \tag{2.10}$$

$A \in [0, \infty)$, is a positive free parameter controlling the flatness of the beta distribution (thus the randomness of the multinomial distribution).

For aggregate data we fit this model by minimizing the negative log likelihood with A determined by minimize function of scipy package from random starting points. We defined a baseline model where all p_k are equal across the rating options (i.e., modeling each rating being randomly selected). We compared the maximum likelihood of the suspense model to the baseline model and average over all individual card draws to present the log likelihood improvement per draw.

2.8.5 Derivation of belief update

In Equation 2.1, we used the relationship $\mathrm{E}_s[\mu^s_{t+1}] = \mu_t$, where s is a game state (i.e. the card score in our context), t is the current time point, μ^s_t is the belief of winning probability given at time t. In other words, the prior belief of winning equals the expectation of the posterior. This is part of the definition of a belief martingale (see Williams (1991), chapter 10) a universal property of Bayesian learning. In Ely et al. 2015 they directly used this property in their definition of suspense.

For an intuitive understanding, consider the following. Imagine an observer has a belief μ concerning the probability of winning a game. We can start from the end of the game then work backwards. Assume under a rule set, there are final states s_w that results in winning and s_l for losing. Thus at the final time step T, the belief in either winning or losing is just determined by the outcome of the game:

$$\mu^s_T = \begin{cases} 0, s \in s_l \\ 1, s \in s_w \end{cases} \tag{2.11}$$

In other words that the final step of the game the belief is just if the person wins or loses. Next, for one time step back at $T-1$ at state s', the winning probability would be marginalizing all the possible next steps weighted by the transition probability:

$$\mu^{s'}_{T-1} = \sum_{s \in s_l} p(s|s') \cdot 0 + \sum_{s \in s_w} p(s|s') \cdot 1 = \sum_s p(s|s')\mu^s_T = \mathrm{E}_s[\mu^s_T] \tag{2.12}$$

Figure 2.12 shows the state transition of the last step, where four possible states are included and marginalized to calculate $\mu^{s'}_{T-1}$. Similarly, the belief μ at any time t, state s is derived from the the belief at time $t+1$:

$$\mu^{s'}_t = \sum_s p(s|s')\mu^s_{t+1} = \mathrm{E}_s[\mu^s_{t+1}] \tag{2.13}$$

Note here we write the term μ_t in Equation 2.1 explicitly as $\mu^{s'}_t$.

2.8.6 Alternative models

2.8.6.1 Formulation of the Heuristic models

First, for the heuristic of "uncertainty", if people keep track of a probability of winning, the 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:

$$S_{\text{uncertainty}} = \mathrm{H}(p_t) \tag{2.14}$$

Alternatively, instead of keeping track of winning probability which requires a full simulation towards the end of the game, people may instead only be concern about how much uncertainty they have now given the 2 candidate cards, or simply, the difference between the two cards:

$$S_{\text{Carddiff}} = |v(1) - v(2)|/(v_{max} - v_{min}) \tag{2.15}$$

where v denotes the value of single card, normalized by the maximum card value difference given all possible values of the cards.

We also tried an alternative normalization using the maxixmum possible difference given the current deck (thus different in every game), but this model does not perform as well.

Secondly, for the heuristic of "fear of losing", if people keep track of a probability of winning, then this can be defined as:

$$S_{\text{pLose}} = 1 - p_t \tag{2.16}$$

In this formulation of pLose model, maximum suspense is achieved when the game is definitely losing. Alternatively, when people are definitely losing, they may instead no longer feel suspense, expressed as almostLose model:

$$S_{\text{almostLose}} = \begin{cases} 1 - p_t, \text{if} p_t > 0 \\ 0, \text{if} p_t = 0 \end{cases} \tag{2.17}$$

Another possible formulation is that people can approximate the closeness to losing by how far is the largest of the two cards drawn from the deck is from the bound:

$$S_{\text{toBound}} = \text{Dist}(\text{Sum}_t + \max(v(1), v(2)) - \text{Bound}) \tag{2.18}$$

Where Sum denotes absolute value, v denotes the value of the next coming card. However, it is very unclear how to normalize this distance. We tried using linear normalization and negative exponential with different constants but none will fit all the games better than the random guess, even after adjusting the constant according to the rules. We deemed this model as an ill-defined model and thus did not report it here.

2.8.6.2 Intuition of the difference between different belief update norms

We here show the magnitude of different belief update metrics given the current probability of winning ($p(now)$) and a potential next step belief ($p(next)$), see Figure 2.13. Note that some combinations are impossible since the expected future belief should be equal to $p(now)$.

More intuitively, we show that when the current belief is 0.5 (i.e. maximally uncertain what

will happen next), the different metrics will give suspense as in Fig2.14.

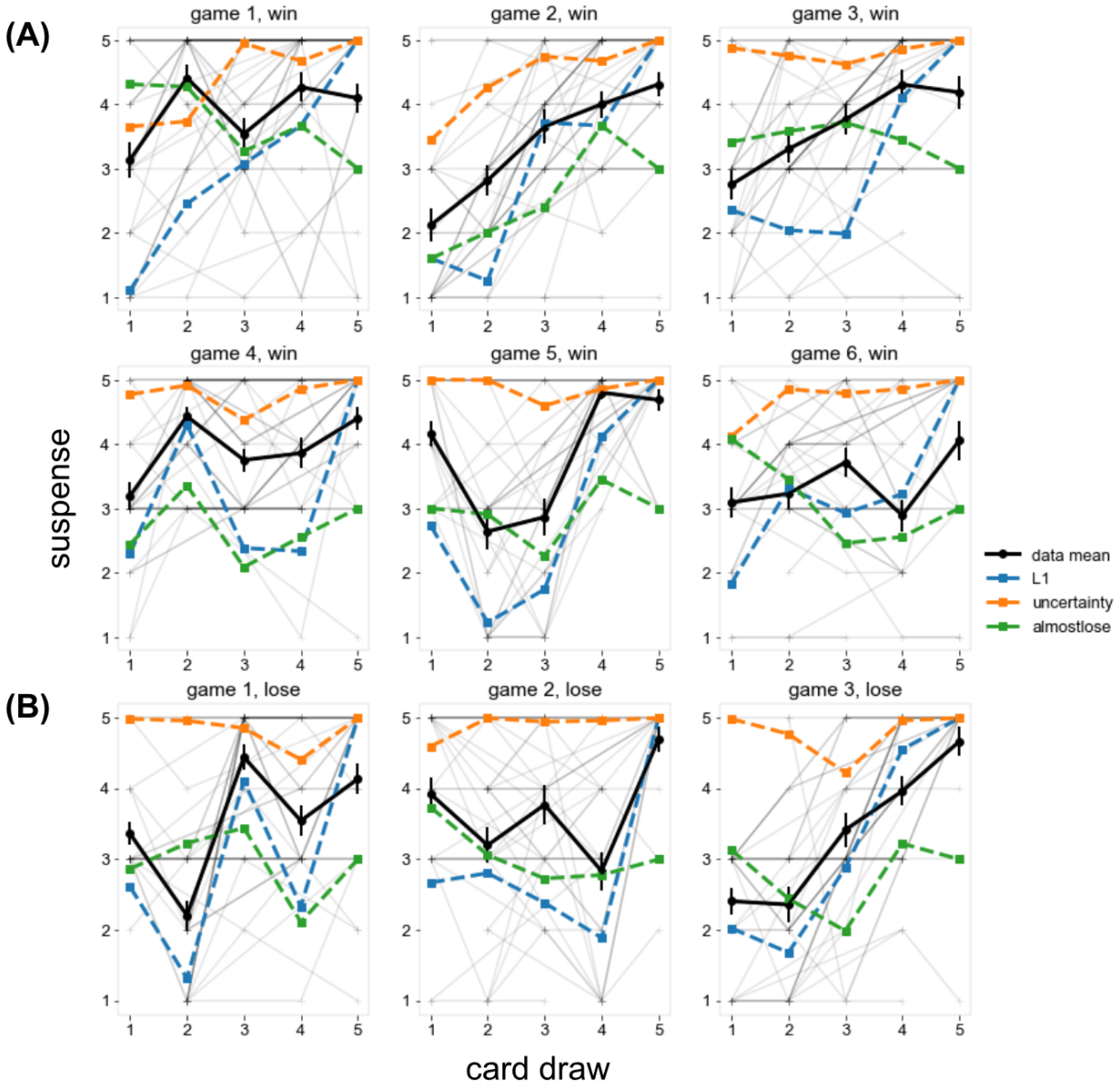

Figure 2.4: Experiment 1: Average suspense judgments and model predictions across all high predicted-suspense games. (A) games resulting in a win; (B) games resulting in a loss

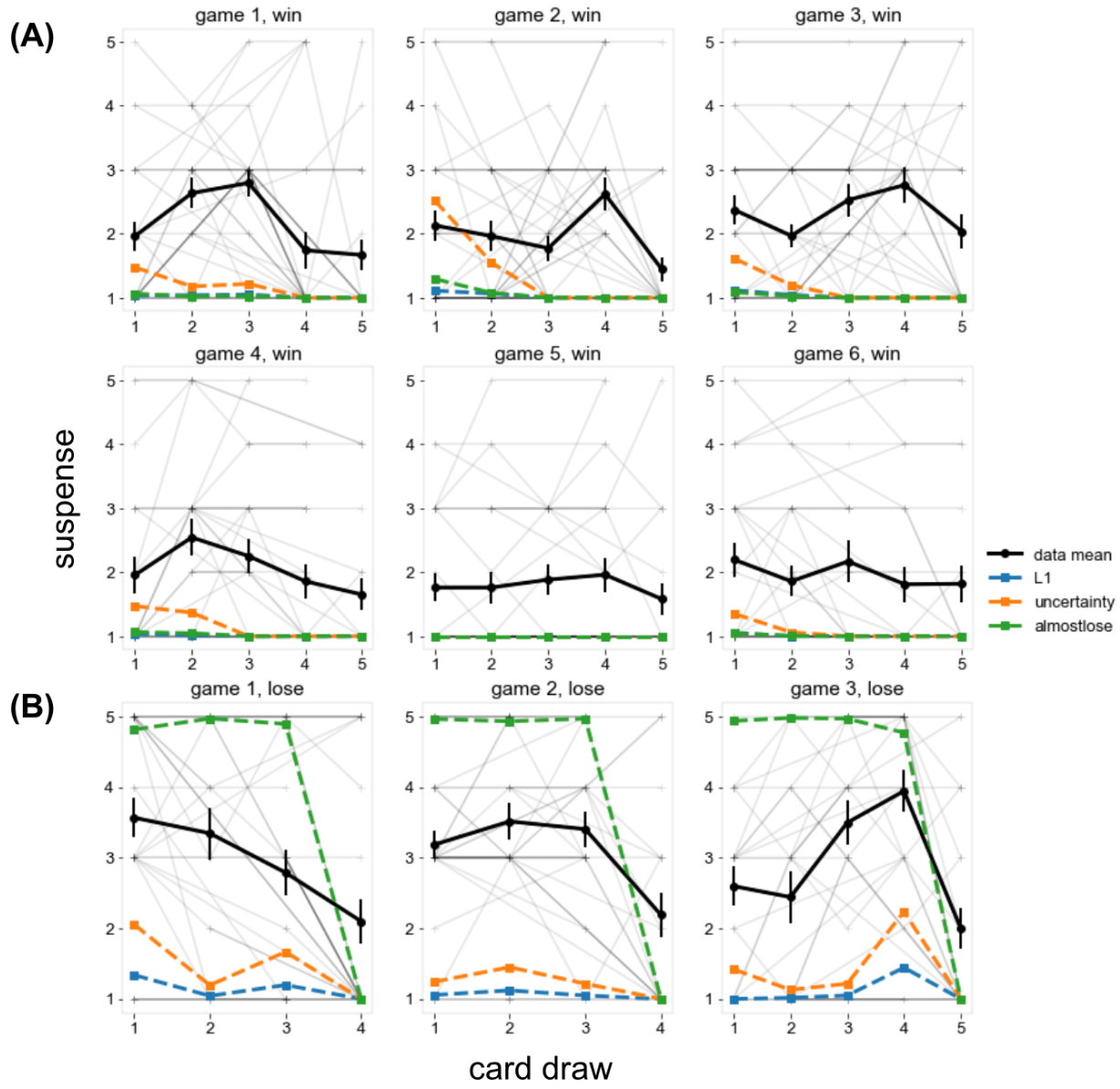

Figure 2.5: Experiment 1: Average suspense judgments and model predictions across all low predicted-suspense games. All the models show systemic biases here. (A) games ended up winning; (B) games ended up losing

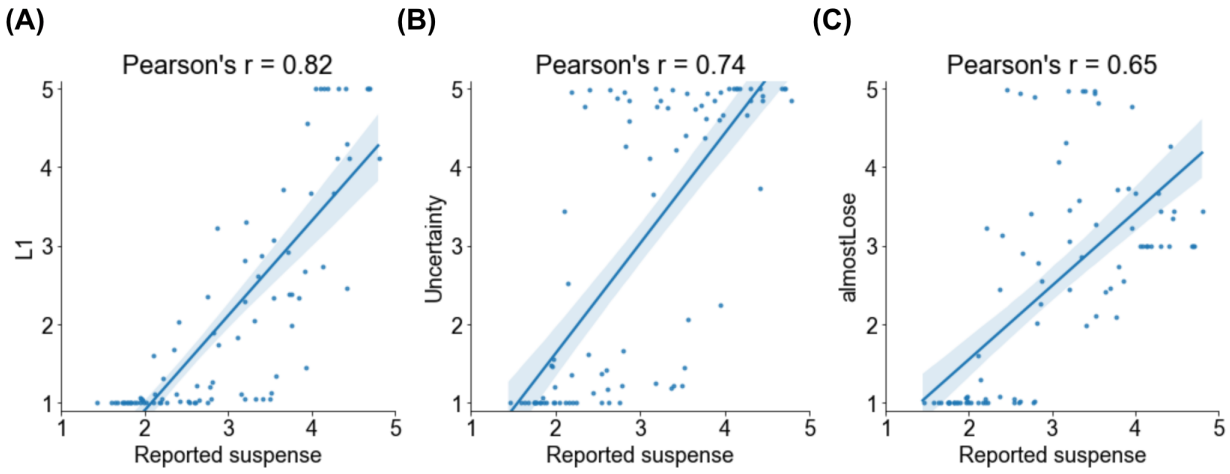

Figure 2.6: Experiment 1: Correlation between model predicted and averaged participant reported suspense. A. "Future belief update" model with L1 norm. B. Uncertainty model C. " Fear of losing" model

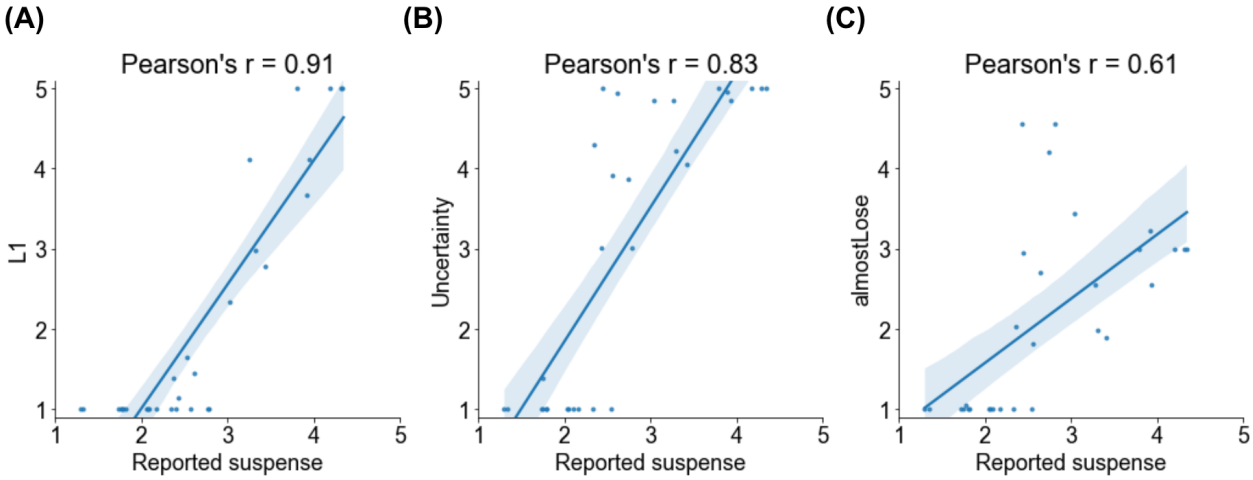

Figure 2.7: Experiment 2: Model predicted and average judged suspense for the best-fitting model from each conceptual class. Games of two rules are all included. The best model from each conceptual category are shown: A. "Future belief update"; L1 norm. B. Uncertainty heuristic; current belief uncertainty. C. "Fear of losing" heuristic; (almostLose)

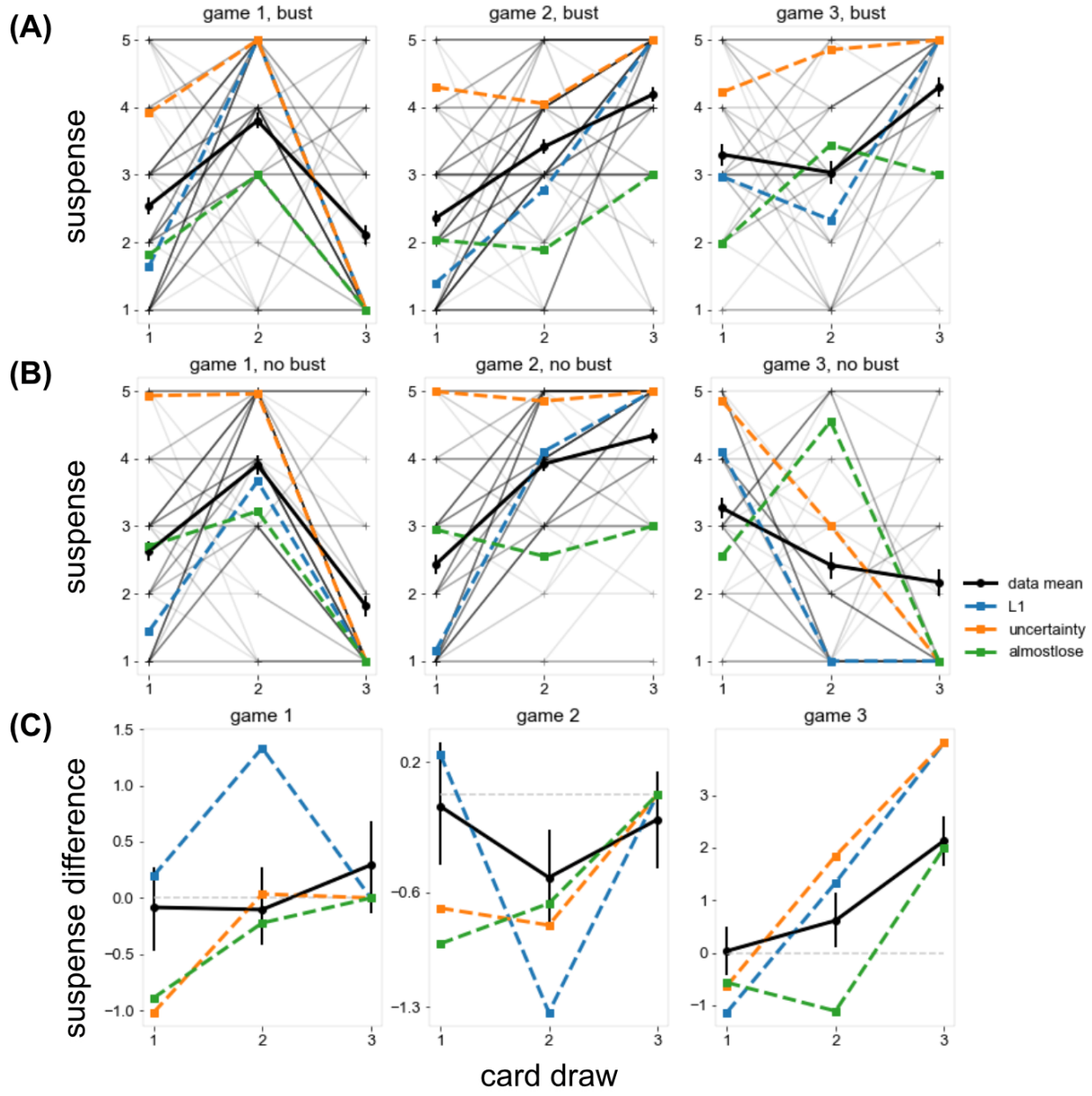

Figure 2.8: Experiment 2: Raw and relative suspense judgments and model predictions across rule conditions for all the games. (A) "Bust" condition. (B) "No-bust" condition. (C) Inter-condition difference ("no-bust" − "bust").

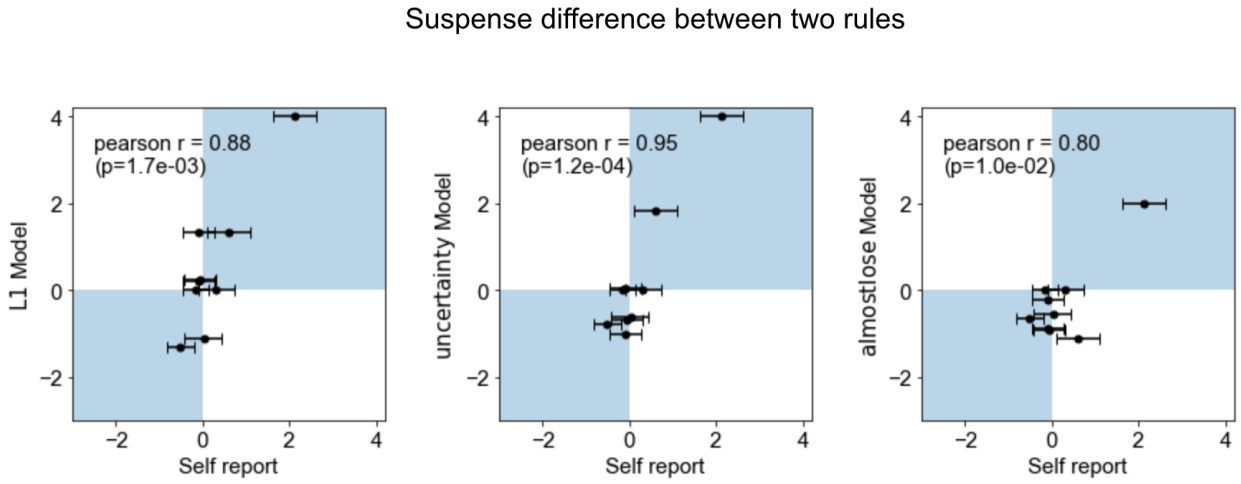

Figure 2.9: Experiment 2, Model predicted and actual suspense difference between "Bust" and "No bust" rule conditions. Each point represents one game point under the two rules. If the model predicts the rule-induced difference to be the same direction of empirical data, the data point will appear in the 1st and 3rd quadrants (shaded in blue). Errorbars on x-axis show bootstrapped 95% confidence intervals (calculated by randomly choosing the ratings between two conditions of the same game point, taking the rating difference and repeating 1000 times. Then the middle 95% of this 1000 samples becomes the confidence interval).

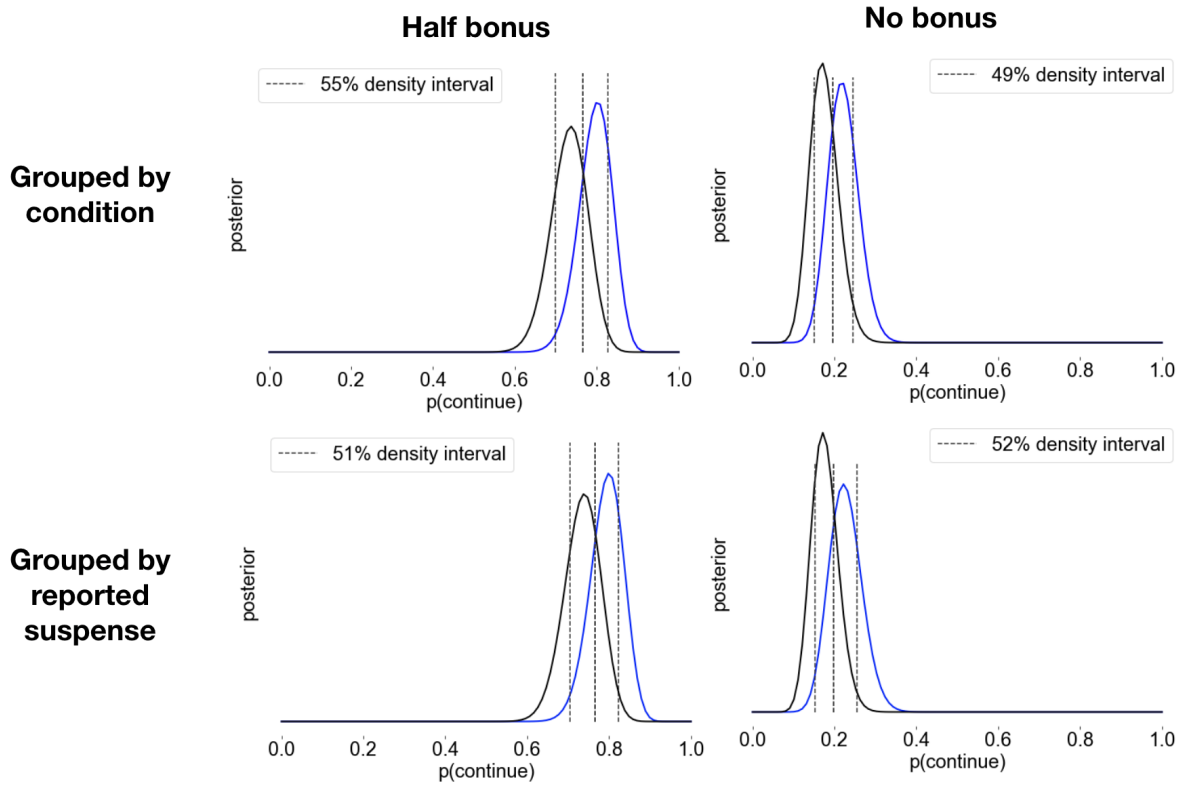

Figure 2.10: Experiment 3, Bayesian posteriors for condition-dependent bias terms. The density interval marks the boundary separating the two distributions. The first row we separated the two distribution by the assigned high/low suspense condition; The second raw we separated participants by whether their self-reported suspense falls into the 50% percentile among the whole sample.

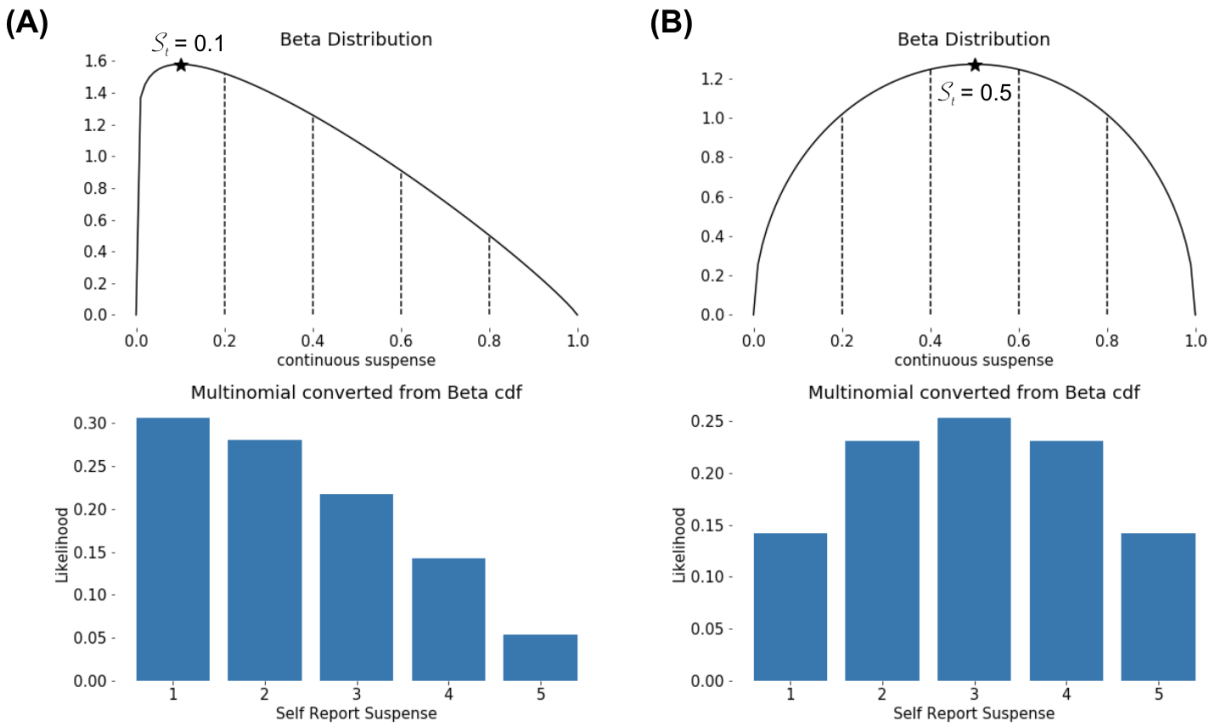

Figure 2.11: Converting the continuous model suspense output to the discrete self report. (A) The model suspense is 0.1, which in turn determines the beta distribution with a mode of 0.1, as marked as the star. Then the area under each 1/5 percentile of this beta distribution (denoted by the break lines) is converted to the value of multinomial for choosing each suspense output. Since model prediction is quite low, in the self reported suspense of 1 being most likely response. (B) With model suspense of 0.5, the the area under each 1/5 percentile of this beta distribution (denoted by the break lines) is symmetric with the highest being the middle percentile. Since model predicts a medium level of suspense, the corresponding multinomial function gives a highest likelihood of responding 3.

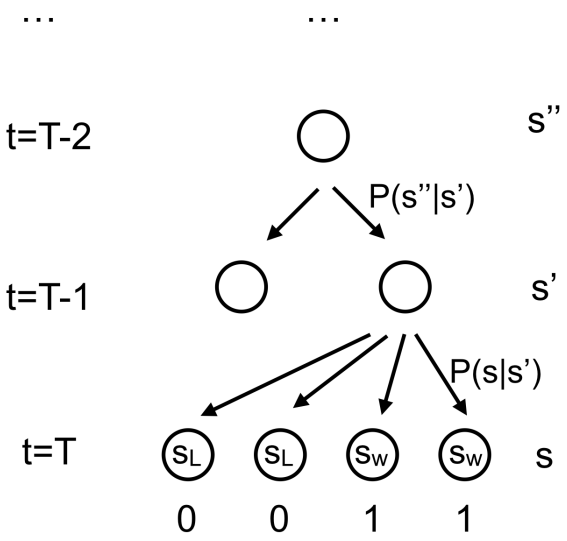

Figure 2.12: Illustration of the last three time steps of state transitions.

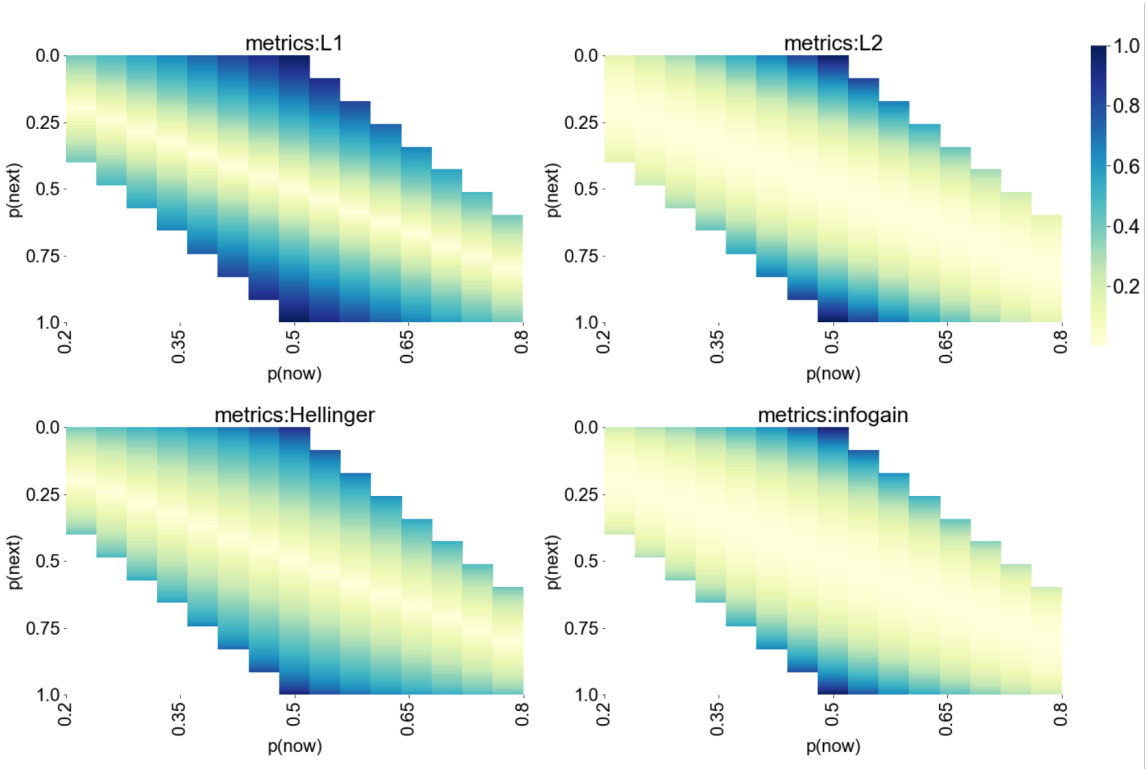

Figure 2.13: Comparing different metrics.

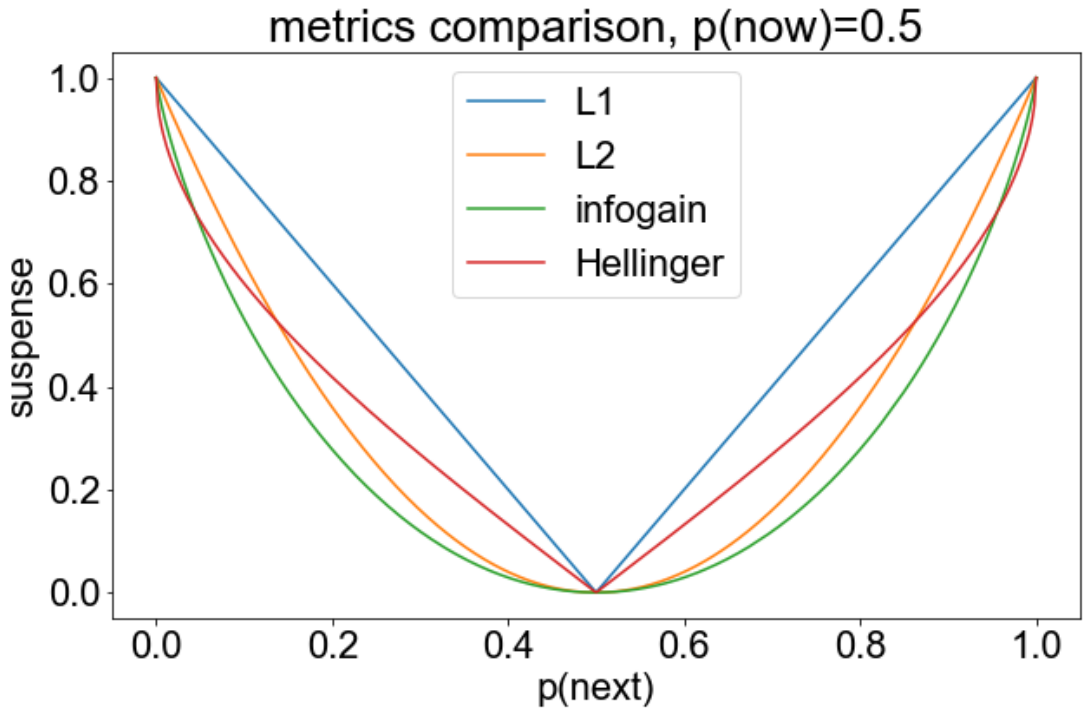

Figure 2.14: Comparing different metrics given the special case of p(now)=0.5