A Quantum Approach to Human Decision Making

Oliver James Waddup

A thesis submitted to the
Department of Psychology
City, University of London
For the degree of
Doctor of Philosophy

June, 2022

Declaration

I grant powers of discretion to the University Librarian to copy this thesis in whole or in part without further reference to the author. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Acknowledgements

Firstly, I would like to offer my sincere thanks to my primary supervisor and co-supervisor, Professor Emmanuel Pothos and Dr. James Yearsley, for their superb supervision and unwavering backing throughout the completion of my PhD. Emmanuel, you have provided the perfect blend of wisdom, encouragement, insight, and humour. Your enthusiastic, 'can-do' attitude fostered nothing but confidence and inspiration in me. James, your ability to explain complex mathematical concepts is extraordinary. I will always be incredibly grateful for your patience. And to you both: I can't thank you enough for granting me the opportunity to pursue the work detailed in this PhD. It has been an honour and a privilege to work alongside you.

I would also like to extend my gratitude to Dr. Pawel Blasiak for the thought-provoking discussions and insightful suggestions that frequently helped turn the mathematically impossible into possible. This PhD wouldn't have been possible without you. A special thanks also to Dr. Bartosz Wojciechowski for his helpful foundational contributions to the Bell Bound work.

The pandemic was a challenging time for all, and I can't thank my "lockdown bubble" enough for keeping me sane through the insane moments. I want to thank Yas Hausler, Aran McKenna, Oliver Grogan, and Chloe Morris for enduring the constant ramblings of a frazzled PhD student in such close proximity for so long. Your patience, curiosity and humour were invaluable to me.

I will always remember the friends and colleagues I gathered whilst working at the Research Lab. Zareen Choudhury and Sharmay Mitchell, thank you for supporting me along my professional journey. Without your guidance, I would not be where I am today.

To my family, you deserve nothing but my endless gratitude. I will be forever grateful for your love and support.

Funding for my work came from the Office of Naval Research Global (grant: N62909-19-1-2000).

Table of Contents

Page

List of Tables

List of Figures

Page

Chapter One: General Introduction

This thesis examines various aspects of decision making, with a focus on probabilistic tools to cognitive modeling. One such tool is the so-called Classical Probability Theory (CPT, or Bayesian Theory; Tenenbaum & Griffiths, 2001; Chater et al., 2006), which has been the dominant approach in understanding different aspects of human behavior. The broad argument is that cognitive processing must reflect some kind of optimal adaptation to environmental statistical structure and, therefore, human cognition must be consistent with the principles of CPT (Oaksford & Chater, 2009). CPT indeed appears to provide accurate descriptions of behavior in many cases, particularly decision making (Siegel et al., 2018), which is the focus of this work.

The challenge to the dominance of CPT in the decision-making literature is in part the result of two of the most influential psychologists of all time; Tversky and Kahneman (the former is one of the most cited psychologists, the latter received a Nobel prize in economics). Tversky and Kahneman provided several examples, in which human decision makers persistently produced judgements in stark contrast with the principles of CPT. For instance, consider their example when participants are asked to decide whether a hypothetical woman, Linda, is more likely to be “a bank teller and a feminist” or “a bank teller”. Because Linda is described as a feminist, but not at all as a bank teller, most participants tend to infer that Prob(bank teller & feminist) > Prob(bank teller). This finding is called the conjunction fallacy (CF, Tversky & Kahneman, 1983). According to CPT (in a single probability space), this is impossible, it is like asking how often it snows and rains in December in London versus how often it just snows in December in London. Clearly, we cannot have more days for the former (conjunctive) event than the latter, this would be impossible.

Modern advances in decision theory have developed and moved beyond CPT as the principal approach to formalising decision making. Whilst, the CF is not compatible with a (basic) CPT framework, we can look to other frameworks, such as Quantum Theory (QT), to consider whether CF decisions can be deemed rational. QT has come to be established as a key alternative formal framework for decision making. In QT, probabilities are computed in a different way (using different axioms) and so the intuitions that emerge for which judgements are appropriate can vary substantially compared to CPT. Indeed, there are many differences between CPT and QT, which provide a nuanced picture for the circumstances when CPT or

QT might be the more appropriate framework for understanding human decision making. For example, in CPT events are definitely true or false, but in QT there are events which can be neither. In CPT, a set of questions can in principle all be resolved concurrently, so that we can talk about the probability of any combination of question outcomes (these joint probabilities always have to exist). In QT, some questions are incompatible, and this means that it is typically impossible to resolve them concurrently. For incompatible questions, certainty for one introduces uncertainty for the other. Probabilistic inference in QT is strongly contextual and perspective dependent, while in CPT it is not (naturally) so.

CPT and QT are both models that allow us to understand the probability of events, albeit in different ways – CPT and QT are based on different axioms and often make different predictions. Let’s explore CPT first. Suppose you rolled a six-sided die. There would be a 1 in 6 chance that you roll a 4. If you rolled a hundred or a million more times, each roll would still offer a 1 in 6 chance of rolling a 4. That is, the probability of the event does not change depending on any subsequent events. Rolling a 4 and then a 6 has the same probability as rolling a 6 and then a 4. This has important implications because, in CPT, any questions we have about our outcomes can in principle all be resolved concurrently. For example, what is the probability of rolling a 4 a hundred times in a row? Indeed, we could talk about the probability of any combination of question outcomes and how these probabilities always exist.

Now let’s consider QT. Suppose now we picked up a new set of six-sided ‘quantum’ die (note, of course, the contrived nature of this example). What would be quantum about them? The outcomes would no longer be able to be resolved concurrently. We would have to use a different basic arithmetic to calculate the probability of a combination of outcomes and the distribution of the die outcomes would be in stark contrast to the one expected from CPT. For example, this time when we roll a 4 and a 6, there is a different probability of rolling a 6 and then a 4. This is because QT is strongly contextual and perspective dependent. This has remarkable implications when we start applying quantum rules to behavioural scenarios. For example, let’s ask someone a set of questions: “do you like your job?” and “are you happy?” Depending on the order in which you answer these questions, you are likely to get very different responses.

CPT is the dominant framework in decision theory, largely due to its descriptive success and ease of understanding. Indeed, Pothos et al. (2021, p. 243) noted that:

> "Whenever an empirical result at odds with classical prescription is identified, the most immediate approach is to explore ways to reconcile it with classical probability theory, then seek explanations based on alternative frameworks. The term 'decision fallacies' exactly refers to findings which are considered hard to reconcile with classical probability theory. A decision fallacy is not just about identifying a result which is inconsistent with classical probability theory. Rather, to characterise a decision as a fallacy it has to be the case that human observers experience persistence in the non-classical intuition, so that even when the classically correct result is explained, they cannot reject the non-classical intuition."

But why even consider QT in the first place? Let's consider several results where we encounter problematic results for CPT, with a view to outline how they can be explained by QT. First, let's suppose we were to answer two questions, A and B. In CPT, the resulting calculation would look like the following: Prob(A)Prob(B|A) = Prob(A&B) = Prob(B&A) = Prob(B)Prob(A|B). In other words, the probability of A occurring and then B is the same as the probability of B occurring and then A. Although this appears intuitive and elegantly simple, there are many occasions when people respond to one question order differently to another, that is: Prob(A&B) ≠ Prob(B&A) (for examples of these question order effects, see Bergus et al., 1998; McKenzie et al., 2002; Moore et al., 2002). QT can predict, or indeed accommodate, order effects whereas CPT assumes commutativity. That is, if we asked two questions, one after the other, such as "are you happy?" and "do you like your job?", in CPT the order in which you ask these questions makes no difference. However, in QT, asking a question (sometimes) inherently disturbs the (mental) state of the individual being asked. Being asked about your happiness second may make you ponder about your happiness in your job versus being asked about your happiness first, whereby you may consider your happiness in all other aspects of your life.

Second, the law of total probability should never be broken. That is, if we consider the probability of an event A in relation to whether X occurs or X does not occur, it should always be the case that Prob(A)=Prob(A&X)+Prob(A&~X). Suppose you want to consider whether, in a prisoner's dilemma task, a participant is likely to defect (D), depending on whether the opponent defects or cooperates. The law of total probability requires that Prob(D) = Prob(D&C) + Prob(D&D), where the second conjunct refers to what the opponent is doing and the first one to what the participant is doing (strictly speaking we should be using indices to differentiate between these events). In words, this states that the probability of the participant defecting must be the sum of the probability of the participant defecting and the opponent cooperating AND the probability of the participant defecting and the opponent defecting. While this constraint makes enormous sense, there are many behavioural situations, which, surprisingly violate the constraint (e.g., Shafir & Tversky, 1992). QT allows for violations of the law of total probability, when the conjuncts 'interfere' with each other and so offers a way to understand the principles which might guide participant behaviour in such cases (Pothos & Busemeyer, 2009).

Third, let us turn to our final example: the conjunction fallacy (CF). Consider a hypothetical person, Linda, who is described as a feminist and not at all as a bank teller (Tversky & Kahneman, 1983). Participants read some information about Linda and are then asked to rank some statements about her (from more likely to less likely). In this example, the key statement relates to whether Linda is a bank teller and a feminist. In most instances, participants indicate that Prob(F&BT) > Prob(BT). But how is it possible to rank the more restrictive category as more likely than the less restrictive one? Psychologists have argued that these findings might result from the contrived nature of the experiment, such that participants might not have fully understood what was being asked of them before they gave their response. However, simpler versions of the CF have been established, such as considering whether a Scandinavian person has blond hair vs. blond hair and blue eyes (Tentori et al, 2004) or when estimating the probability of weather events in a particular place (Costello & Watts, 2014). In QT, these findings can be written as Prob(BT) = Prob(F & then BT) + Prob(~ F & then BT) + Δ, an expression which resembles the law of total probability plus Δ. The latter is an interference term which allows violations of the law of total probability in general and, specifically, it can allow Prob(BT) < Prob(F & then BT). Note, for incompatible questions we have assumed that conjunctions are meaningful only in a sequential form, that is, instead of Prob(F & BT), we have to write Prob(F & then BT); in the latter term we make an explicit assumption about the specific order in which the questions are assessed.

To summarise, these are some examples where results paradoxical from a classical perspective make sense using quantum principles. This is the general contribution from quantum theory in psychology: it has provided a formal framework which has helped us understand some ‘paradoxes’, especially in decision making.

Before we delve any deeper into this work, we wanted to highlight two unrelated but important conceptual and practical details surrounding the preparation of this thesis. Firstly, one might wonder why a Psychology PhD student decided to write a thesis on reconciling human decision making with ideas known only to those who study theoretical physics. This work was funded by the Office of Naval Research Global to form an interdisciplinary team of mathematicians and psychologists to begin to resolve how concepts, such as the Bell inequality, can be ‘translated’ into psychological, “tangible” study. For example, in physics particles elicit ‘spooky action at a distance’: one particle can ‘know’ something about another particle even if separated by immense distances. In real terms, if a person does not pick up the telephone (or through some other form of communication) to tell their friend on the opposite side of the world how their coin flip landed, their friend would simply guess the outcome. So how do we reconcile quantum effects with human participants? In this thesis, we offer some practical ideas towards resolving questions such as these.

Secondly, much of the experimental work was completed during the COVID-19 pandemic. Importantly, this meant that all lab-based experiments were prohibited for the health and wellbeing of all involved (the participants, experimenters, as well as other staff, students, and workers on the university campus). Our experiments were designed with this in mind, and so it was essential that all experiments could be conducted online. Of course, we had planned for several lab-based projects in the early months of this work, but it was not feasible to carry them out. For example, as will be seen in Chapter 3, we attempted to replicate lab-based work, but we had to refine it to be an entirely online protocol. The extent of the impact this had on our results is not known.

Let us now return to the thesis at hand by exploring the structure and content of this work. This thesis is presented in five parts. The current chapter serves as a general introduction to the current work. Chapters 2 through 4 present experimental studies testing the utility of QT in different decision-making contexts. Chapter 5 presents as a general conclusion, outlining the theoretical gains and limitations of this work.

In Chapter 2, we used Prisoner's Dilemma (PD) games to simulate interactions between two people. Suppose we have two players, Alice and Bob, who each have two binary questions. In PD games, it is common for players to be asked to co-operate or defect (e.g., for Alice we have $a1$, $a2$ and for Bob we have $b1$, $b2$, each having two possible outcomes). One might reasonably expect that it is possible to represent probabilities from such tasks as marginals from a four-way joint probability distribution. So, our question is, are there occurrences in PD games for which participant behavior might diverge from this expectation? In this chapter, we test whether the outcomes of these tasks can be illustrated beyond a four-way probability distribution. Specifically, we were interested in whether there was a *sensitivity to context*, which can briefly be defined as responses to a question that depend on the other person's questions. That is, perhaps certainty for one of Alice's outcomes creates uncertainty for Bob's. The implication of sensitivity to context is that it may preclude the existence of a joint probability distribution and also reveal so-called supercorrelations – correlations 'stronger' (in some specific sense) than the strongest possible standard correlations.

We explored this research question using the probabilistic tools mentioned above, as well as the idea of Bell's bound, a threshold value that allows us to determine whether CPT or QT is more predictive of the observed data (specifically, we define a quantity S: the sum of probabilities of action across both questions for both players. Briefly, when $S \leq 2$, outcomes are bounded by CPT, and when $S > 2$, properties of QT may be present. All this is presented in detail in Chapter 2). We conducted a series of five experiments, each with improved methodology. The final two experiments were more refined and were the only experiments included in our modeling. Results revealed that participants were sensitive to the context in the PD games, and our observed S values exceeded Bell's bound, broadly in line with predictions made by the QT model. Moreover, fits by a classical model were in line with expectation (bounded by $S=2$), further adding to the body of research that QT sometimes offers a better descriptive framework for behavior than CPT (Busemeyer & Bruza, 2011; Pothos et al., 2013).

A crucial aspect of QT is that an action can change the system. Analogously, a participant's decision, choice or evaluation can change their psychological state (Sharot et al., 2010; White et al., 2014). In Chapter 3, we develop this idea further. Specifically, we explore constructive influences and how future judgements can be shaped by the impressions and evaluations that occur before them. For example, if a person states they like a particular food, does their preference for this food change? Intuitively, one might expect that responses follow an existing judgement or attitude, that is, if you state that you like chocolate, this reveals your pre-existing attitude towards chocolate. In our work, we explore an alternative argument: that expressing an impression influences how you judge the same stimulus in the future, contingent on how you have interacted with the stimulus in the past (qualified shortly). This argument is not novel (e.g., see Sharot et al., 2010), but our exploration using tools from QT is.

White et al. (2014) created a model which predicted constructive effects for affective evaluations on advertisements. They examined whether the expression of an opinion on a positively or negatively valenced advertisement would impact future judgements of another advertisement. They found that if participants rated a negative advert, involving a burning building, first and then a positive advert, with a loving family enjoying a meal together, the negative advert would be rated more positively than if it was considered by itself. Similarly, an evaluation of a positive advert and then a negative advert would reveal more negative impressions than if the positive advert was evaluated by itself. White et al. (2014) labelled this effect the 'Evaluation Bias' (EB).

Research has demonstrated the EB using different contexts, including judgements of trustworthiness and questions regarding the state and strategy of the participant's employer (White et al., 2016; 2020). The present chapter also aimed to replicate and extend the work of White et al. (2014). We sought to identify the conditions required for an individual to show the EB in their judgements. We examined the role of self-reflection in individuals and how this can help predict the EB. Specifically, we focused on two areas of self-reflection: metacognition and mindfulness. These two self-reflective processes are both important and we used these to examine the extent to which the EB may be related to an individual's awareness of their decisions (Brown & Ryan, 2003; Dunlosky & Metcalfe, 2008).

We first aimed to replicate White et al. (2014). To do this, we conducted a pilot experiment to create baseline valence measurements of our advertisements. This allowed us to remove stimuli from the experiment that did not fit with their intended valence and enabled us to compare our ratings to those of White et al. (2014). After filtering the stimuli, we then conducted three experiments. The first experiment suffered from a stimulus randomisation issue that was resolved for the subsequent studies. In terms of participant recruitment, Experiments 1 and 2 were conducted using Prolific and Experiment 3 more closely replicated White et al.'s protocol by using sampling from a university population. Our results indicated that all three experiments were unable to replicate White et al.'s (2014) EB. We also found no evidence that mindfulness and metacognition impacted propensity for the EB. Instead, we found equivalent ratings between our positive and negative adverts (and vice versa). Because of these results, we turned to an alternative approach which closely examined the valence ratings for when the adverts were shown first relative to when they were shown second. One notable finding was that negative adverts impacted positive adverts more so than the other way around. The implications of this result are further discussed in Chapter 3.

In Chapter 4, we examined the Temporal Bell (TB) inequality in human decision making, a mathematical test of QT characteristics in behavior (Leggett & Garg, 1985; Atmanspacher & Filk, 2010). Instead of studying the correlations of actions across two people and two binary questions (the Bell Bound), in Chapter 4 we focused on the correlation values generated from an individual's single question asked across three different pairs of time points (12, 23 and 13). The TB inequality can be shown as:

$$C_{12} + C_{23} \leq C_{13} + 1$$

where the sum of the correlations between times points 1,2 and 2,3 are bounded by time point 1,3 + 1 (analogous to standard Bell Bound of 2). Before we explain the meaning of violating TB, it would be helpful to first explain its assumptions.

There are three assumptions of the TB inequality: macrorealism, non-invasive measurability (NIM), and the arrow of time. Note, we will only offer a brief summary of the assumptions here, but we will further explore them in Chapter 4. Firstly, macrorealism implies that a system is always at a specific state. Suppose a jury member was asked to render a verdict on a suspect after every piece of evidence was shown to them. Macrorealism means that the juror would definitively consider the suspect to be guilty or innocent every time they were asked the question. Secondly, NIM relates to whether the act of measuring (or asking the juror for their verdict, as in the above example) impacts the outcome. Thirdly, the arrow of time assumption simply means that earlier events can influence what may occur later, but not vice versa.

Taking these assumptions together, the goal of a TB inequality test is to test for quantum character in a system. Violating TB suggests that one of the assumptions does not hold. For instance, let us assume macrorealism is violated. This would mean that a juror's verdict is not definitively guilty or innocent until they are asked. In other words, jurors construct their verdict at each different time point. If TB is not violated, then all its assumptions are satisfied, and thus we can assume a classical trajectory of the verdicts. Specifically, this means that at all times a decision maker would consider the suspect as definitely guilty or definitely innocent or a (definite) mixture of the two (by contrast, quantum-like uncertainty means that when trying to resolve uncertainty you cannot know with precision whether the judgement will go towards e.g., guilt or innocence). That is, the conditions of the assumptions above will hold.

We utilised a hypothetical murder trial (Yearsley & Pothos, 2016) to provide a simple demonstration of how QT can violate the TB inequality. Participants assumed the role of a juror and were shown some evidence across different days before being asked whether their verdict had changed across one of the three pairs of time points. Note, asking whether participants had changed their mind between the different time points allowed us to measure change in opinion between time points, without having to ask two separate questions. This distinction relates to the assumption of NIM. That is, if we measure too coarsely the verdicts this can impact future queries of the juror's response (see the Quantum Zeno effect in Yearsley & Pothos, 2016). Subsequently, we wanted to ensure that our 'change' judgements were not the result of two individual judgements and a computation of the difference. Our solution to this problem was to record response times (RTs) for when participants responded to marginal questions as well as when they responded to the change questions. We reason that if the RTs are different then participants can create the change judgement independently from the marginals.

We conducted five experiments, as well as several pilot and control experiments to refine the paradigm between iterations. Results showed we were unable to violate TB in the first few iterations of the paradigm. The supplementary pilot and control experiments allowed us to test several key aspects of the study, such as adjusting scenarios and variations of evidence. Note, concerning this latter point, we carried out some approximate simulations on the strength of the evidence, specifically indicating which evidence pieces would most likely lead to a TB violation. It was not until Experiment 4 that we were able to violate the TB inequality. We later supported this finding with unequal RTs in Experiment 5. A TB violation suggests that a quantum-like structure was (in general) present in this decision paradigm. Whilst there was a gradual methodological improvement across experiments, we should also be sceptical when events occur under highly prescribed conditions. Subsequently, future research in this area should be aimed at replicating our results and extending the work to include other contexts.

Finally, in Chapter 5 we provide an overview of the results from this thesis. We revisit Chapters 2 through 4 and explore the implications of our findings in relation to CPT and QT. For instance, in Chapter 2 we provided the first demonstration of the Bell framework into two interacting people. Our results showed that participants were sensitive to the context we created, and the empirical S values exceeded Bell's bound. To some this may not be surprising, given the highly controlled conditions of the experiment. However, the more surprising implication is that this sensitivity prevented fits by a simple classical model and so shows another way in which PD tasks can produce results problematic for baseline expectation from CPT. In Chapter 3, we were unsuccessful in replicating White et al.'s (2014) EB, and we also failed to produce any mediating effects of introspection. Reasons for these failures might relate to the stimuli used (we employed different adverts to White et al.'s pilot, and these adverts were stronger in intensity) or conceptual misunderstandings (our failure to anticipate the implications of different types of mindfulness), to name a couple. However, as mentioned above, after an alternative approach, we did lend some support to other research in an unexpected direction. Notably, we found that initially negative information seems to have a higher impact on the perception of the second positive stimulus (than vice versa). Chapter 4 was the most conceptually challenging chapter in this thesis. It was fraught with technical details (such as the assumptions of TB and others such as non-signalling in time and equal signalling in time that we explain further in this thesis) that all seemingly served to halt our progress. It took several iterations of the paradigm before we were able to violate the TB inequality. This violation of the TB inequality can be taken to be evidence against macrorealism, that is, the assumption that a question can have a specific value at all times. Indeed, the implications of this result mean that, without macrorealism, how one constructs mental representations (such as a juror verdict) will be impacted by order or interference effects. To conclude our final chapter, we will also highlight some future directions for research.

Chapter Two: The Bell Bound

The content of this chapter is mostly taken from our relevant publication (Waddup et al., 2021; see statement of co-authors of joint publication, p.236). In our published work, we provide the same rationale and motivations for the experiments, but we examine only two experiments (notably, Experiments 4 and 5 outlined below). In this chapter, we offer the full account of our experimental paradigm, including all experimental iterations and the discussion of our results with respect to classical and quantum models.

2.1 Abstract

Considering two agents responding to two (binary) questions each, we define *sensitivity to context* as a state of affairs such that responses to a question depend on the other agent's questions, with the implication that it is not possible to represent the corresponding probabilities with a four-way probability distribution. We report five experiments with a variant of a PD task (but without a Nash equilibrium), which examine the sensitivity of participants to context. The empirical results indicate sensitivity to context and add to the body of evidence that PD tasks can be constructed so that behavior appears inconsistent with baseline CPT (and the assumption that decisions are described by random variables revealing pre-existing values). We fitted two closely matched models to the results, a classical one and a quantum one, and observed superior fits for the latter. Thus, in this case, sensitivity to context goes hand in hand with (epiphenomenal) entanglement, the key characteristic of the quantum model.

2.2 Introduction and Basic Definitions

PD games involve two players with a binary action each, typically denoted as cooperate (C) versus defect (D). A usually symmetrical payoff matrix determines the reward of each player, depending on their combined action. Typically, payoffs are set so that it is most advantageous to D, if the other player Cs, but the mutual gain is highest if they both C (defection is then the Nash equilibrium). PD games have been extensively studied in psychology, partly because they can lead to apparent discrepancies with CPT (Broekaert et al., 2020; Chater et al., 2008; Pothos & Busemeyer, 2009; Shafir & Tversky, 1992). In the pioneering study by Shafir and Tversky (1992), participants were put in the shoes of one of the players in a PD game and were presented with three kinds of trials: first, trials for which participants were told the other player would defect; second, trials for which participants were told the other player would cooperate; third, trials for which participants were not given information about the other player. Results indicated that $Prob(D_{Participant}, unknown)$ was outside the bounds of $Prob(D_{Participant}|known\ C)$ and $Prob(D_{Participant}|known\ D)$, thus violating the law of total probability. Such results are not insurmountably inconsistent with CPT, but they do challenge the ubiquitousness of CPT in cognitive theory (Griffiths et al., 2010; Khrennikov, 2004; Oaksford & Chater, 2007; Tenenbaum et al., 2011).

In standard PD paradigms, there is a Nash equilibrium for each participant to D, that is, neither participant can improve her position by unilaterally changing a D action. In this work, we do not consider such PD paradigms, but rather just the two-player interactions, based on a payoff matrix without a Nash equilibrium. We refer to such paradigms as PD variants. The surprising hypothesis we are interested in is whether there are PD variants for which choice statistics cannot be modelled with a four-way probability distribution (this statement will be qualified shortly). So, our paradigm reflects a minimal set up of interaction between two agents. While there is a vast literature on game theory, we avoid engaging with this literature so as to focus on our specific objective: are there simple situations for the interaction between two agents, as just described, which might confound the straightforward expectation that behavior can be modelled with a four-way probability distribution?

Consider a PD variant, such that each of two players, Alice and Bob, have two binary questions; Alice's questions are $a1$, $a2$ and Bob's $b1$, $b2$, all having two possible outcomes ± 1. A baseline classical expectation is that it is always possible to represent probabilities from such tasks as marginals from a four-way joint probability distribution. More conventionally, we expect that corresponding choice frequencies can be organized in a four-way table. So, our question is, are there PD variants for which participant behavior might be inconsistent with this expectation?

Noting that expectation values are computed as $\sum_{all\ possible\ outcomes} Prob\ (outcome_i) \cdot Value(outcome_i)$, for a pair of binary questions, x, y, with $Prob\ (outcome_i)$ being the probability of $outcome_i$ and $Value(outcome_i)$ the value assigned to $outcome_i$, the expectation value is

$\langle x\&y\rangle = Prob(+ + |x, y) \cdot 1 \cdot 1 + \ Prob(- - |x, y) \cdot (-1) \cdot (-1) + Prob(+ - |x, y) \cdot 1 \cdot (-1) + Prob(- + |x, y) \cdot (-1) \cdot 1$ ……………………………….…………Equation (1)

Define the quantity:

$S = |\langle a1\&b1\rangle + \langle a1\&b2\rangle + \langle a2\&b1\rangle - \langle a2\&b2\rangle|$………………………………Equation (2)

Consider three conditions when computing these expectations. First, locality means that Alice answers her questions without any information about what Bob is doing, and vice versa. Locality means that Alice and Bob are separated in space and no communication between them is possible (Atmanspacher & Filk, 2019). Second, free choice means that the question asked to Alice is determined independently from the one asked to Bob. Third, realism means that the outcomes to Alice and Bob's questions exist, whether Alice and Bob state them or not. One of the most significant results in theoretical physics is that, with locality, free choice, and realism, the maximum value of *S* is 2; this upper limit of *S* is called Bell's bound (Bell, 1964, 1987; Clauser et al., 1969). Let us take realism for granted, so henceforth we will focus on locality and free choice.

Note, locality and free choice are properties of the two systems producing the relevant statistics. So, in the example with Alice and Bob, locality means that the two agents are local relative to each other – there is no communication – so that Alice has no information about Bob when making her choices and vice versa. Likewise, free choice means that Alice's choices are not influenced by Bob's.

How can Bell's bound be broken? Consider Alice and Bob perfectly tuned to each other, so that $\langle a1\&b1\rangle = 1$, $\langle a1\&b2\rangle = 1$, and $\langle a2\&b1\rangle = 1$. Given this, if locality and free choice apply, then questions $a2$, $b2$ must correlate as well. This is because if $a1$, $b1$ perfectly correlate and $a1$, $b2$ perfectly correlate, then $b1$, $b2$ must perfectly correlate too. This, together with the fact that $a2$, $b1$ perfectly correlate with each other, leads to the conclusion that $a2$, $b2$ must perfectly correlate as well. But, if $\langle a2\&b2\rangle = 1$, then the S=2, which is the maximum value that S can take, with realism, locality, and free choice. Therefore, the only way we can break Bell's bound is via some kind of *sensitivity to context*. For example, Bob's answers are sensitive to the context created by Alice's questions.

To explain sensitivity to context, suppose that the $b2$ question depends on whether Alice considers $a1$ or $a2$. If Alice considers $a1$, then Bob responds to $b2$ in a way that the two questions correlate with each other, $\langle a1\&b2\rangle = 1$. However, if Alice considers the $a2$ question, then Bob responds to $b2$ in a way that the outcomes of the two questions anticorrelate, $\langle a2\&b2\rangle = -1$. That is, there is no answer to the $b2$ question, independently of what Alice does. If we accept the possibility of sensitivity to context, then we can easily see that the Bell bound can be exceeded, in that $S = |\langle a1\&b1\rangle + \langle a1\&b2\rangle + \langle a2\&b1\rangle - \langle a2\&b2\rangle| = 1 + 1 + 1 - -1 = 4$. In this simple situation, sensitivity to context means that the original set of questions $\{a1, a2, b1, b2\}$ is better understood as $\{a1, a2, b1, b2_{a1}, b2_{a2}\}$, where $b2$ has two different versions, depending on which question Alice responds to.

Cases when S>2 reveal a case of correlation 'stronger' than classical correlation. For S>2, it is not sufficient for pairs of questions to be responded perfectly in tune with each other (this would be a case of perfect, classical correlation). It is also required that responses are sensitive to the questions asked by the other agent. Thus, cases of S>2 can be said to reflect *supercorrelation* (noting of course that correlation is a binary relation, whereas supercorrelation is a relation between answers amongst two sets of questions). As noted, in the physics literature, the kind of correlation producing S=4 is called a PR-box and refers to the strongest type of non-local correlation that is non-signaling, in the two-question, two-outcome scenario (Popescu & Rohrlich, 1994).

Especially in physics, this discussion is complicated by various inter-related notions, such as signaling, disturbance and communication. Signaling is a statistical notion informing us of whether the choice of measurement on one side effects the statistics on the other side. The idea is that Alice and Bob have some device generating statistics $Prob(ab|xy)$, where a, b indicate outcomes and x, y=1,2 are the measurement settings for Alice and Bob respectively (measurement settings mean which questions are asked). Signalling is if Alice is able to send a meaningful signal to Bob concerning what her setting is, x=1,2. If signalling occurs, then Bob can infer Alice's measurement setting by looking at the statistics on his side, i.e., depending on whether his statistics are different for different measurement settings for Alice: $Prob(b|1y) \neq Prob(b|2y)$. Let us note that, if Bob does not know the outcome of Alice's measurement, then we have to marginalise across different possibilities for this outcome, writing, e.g., when we are interested in x=1, $Prob(b|1y) = \sum_{a=+1,-1} Prob(ab|1y)$. So, the signaling condition is that $Prob(b|1y) = \sum_a Prob(ab|1y) \neq \sum_a Prob(ab|2y) = Prob(b|2y)$, that is, as noted, that Bob can tell whether Alice measures x=1 or x=2, by looking at the statistics on his side (later on, in the Signaling section we offer an equivalent way to compute signaling quantifiers).

When there is no signaling, another seminal result, Fine's (1982) theorem, shows that one condition for the existence of a (four-way) joint probability distribution for four binary random variables is $S \leq 2$, which is called the Clauser, Horne, Shimony, and Holt (CSHS) inequality (Clauser et al., 1969). Note, there are four versions to the inequality, depending on which expectation is given a minus sign in Equation 2 and Fine's result states that the bound 2 for all those four expressions is the sufficient condition for the existence of a joint probability distribution. When there is signaling, there is a corresponding generalized test of contextuality due to Dzhafarov et al. (2016; but see also Atmanspacher & Filk, 2019). Above we referred to sensitivity to context rather than contextuality. We will define sensitivity to context more precisely shortly and offer our rationale for why sensitivity to context is the more appropriate notion for the present work, as opposed to contextuality. Readers should note, however, that there is intense, ongoing debate on these issues.

Presently, what we are interested in is whether there is sensitivity to context, which can be defined as the non-existence of a joint probability distribution – informally, we can say that Alice changes her answer to her question, depending on the question that Bob has. When there is signaling, we can immediately conclude that there is sensitivity to context, regardless of whether $S > 2$ or $S \leq 2$. However, sometimes we may want to test for sensitivity to context without considering signaling. For example, this might be because signaling is low and hence our estimate of signaling is not necessarily reliable (for an example in physics, see Adenier & Khrennikov, 2017). In such cases, when $S > 2$, we can conclude that there is sensitivity to context (this follows from the usual proof of the Bell inequalities based only on the factorization property for conditional probabilities).

Here is the tricky point: Dzhafarov et al.'s (2016) generalized test examines sensitivity to context when there is signaling (their expression can be seen as subtracting away the influence from signaling). But in the present case, the only interest is whether Alice employs the available information of what Bob does, to demonstrate sensitivity to context (here and throughout as defined in the paragraph above), regardless of whether this is due to signaling or not. So Dzhafarov et al.'s (2016) generalized test is not relevant here.

These distinctions are particularly relevant in psychology, since the only systems known to break Bell's bound are physical systems of microscopic particles, obeying the laws of quantum mechanics. By contrast, for macroscopic systems, it is generally (see shortly) accepted that violations of Bell's bound can be accounted for only by communication, disturbance or some other equivalent mechanism, between the two systems (Atmanspacher & Filk, 2019). For example, demonstrably classical systems, such as containers with fluids at different levels, connected by tubes allow the construction of variables which violate Bell's bound. But, of course, there is nothing peculiar going on and this is just a result of communication or influence between the systems (such examples have been known for a while, e.g., Aerts, 1982, or Toner & Bacon, 2003). We can say that such systems demonstrate sensitivity to context. Note, there are subtleties to this discussion, for example see S. Aerts (2005; also D. Aerts, 2014), who described possible systems for which a measurement (decision) itself can bring about the dependence to context needed for S>2. An additional subtlety is whether communication is assumed to lead to signaling or not. In Toner and Bacon (2003) and S. Aerts (2005) there is no signaling, but in D. Aerts (1982) there is signaling (as Toner & Bacon note, *in general*, communication can be taken to be some influence of some sort, but it does not always have to lead to signaling). These ideas are interesting, though we think they do not apply to the present results (this issue is briefly considered in the General Discussion).

2.3 Psychological Implications and Outline

Bell’s bound has an almost magical quality. Sensitivity to context means impossibility of describing the system in the usual way via a four-way probability distribution, with the marginal distributions representing the observed (conditional) statistics. But what exactly does this mean? Consider Table 2.1, wherein we assume that all marginal probabilities are 0.5. For the right-hand side, *S*=4 and it can be shown that the corresponding probability information is not self-consistent (the same conjunction can be ‘shown’ to be both zero and non-zero). We think that, amongst experimental psychologists at least, it is a baseline expectation that probabilities can be organized in a table of this kind.

Table 2.1. Proportions of +, - responses for a PD variant. Note, each table is four separate probability subtables, corresponding to different measurements for the two systems. For the left table *S*=2 and for the right *S*=4>2. It can be shown that the right table is inconsistent. The right table is a famous one, corresponding to the Popescu-Rohlich box (PR-box; Popescu & Rohrlich, 1994).

	*b*1=+	*b1*=-	*b*2=+	*B*2=-		*b*1=+	*b*1=-	*b*2=+	*b*2=-
*a*1=+	0.5		0.5		*a*1=+	0.5		0.5	
*a*1=-		0.5		0.5	*a*1=-		0.5		0.5
*a*2=+	0.5		0.5		*a*2=+	0.5			0.5
*a*2=-		0.5		0.5	*a*2=-		0.5	0.5	

We are interested in how these ideas translate to two individuals playing a game, corresponding to a Bell scenario (i.e., each individual has two binary questions). Of course, an interaction between two individuals is an extremely common decision situation. With the locality and free choice assumptions, in general it is impossible to break Bell’s bound (Bell, 1987; cf. S. Aerts, 2005, and D. Aerts, 2014). For two agents, the only way Bell’s bound can be exceeded is if at least one of the free choice or locality assumptions is violated. For example, suppose we retain free choice and allow violations of locality. Then, Bob needs to adjust his answers depending on knowledge of which question Alice receives. So, the decision to stay local or not is ‘outsourced’ to Bob – in the experimental paradigm we employ, it is up to the participants (on a trial-by-trial basis) to decide whether to stay local or not. This is the essence of the paradigm we will shortly present.

So far, while there have been several studies concerning Bell's bound in psychology, these studies have focused on the thought processes of individual participants. Specifically, there have been several examinations of sensitivity to context, for the same participant answering all four questions, $a1$, $a2$, $b1$, $b2$ (for an early example see Conte et al., 2008). The issue of compositionality in conceptual combination concerns whether the constituent concepts combine in a way that their meaning independently determines the meaning of the composite concept. For example, in considering the novel conceptual combination 'spring plant', under a compositionality assumption we would look for some meaning from 'spring' and some from 'plant', independently combined together. A contrasting hypothesis is that a constituent in a conceptual combination acquires meaning contextually, depending on the other constituent. For example, in the case of boxer-bat, whether we consider a sporting or animal sense for 'bat', will impact on the how we interpret 'boxer' (Bruza et al., 2015). A number of theorists have employed the CHSH inequality or variants to conclude in favor of non-compositionality in conceptual combination (Bruza et al., 2015; cf. Aerts et al., 2016), an issue of considerable significance concerning conceptual representation (see also Hampton, 1988; Osherson & Smith, 1981; Storms et al., 1999). Similar ideas have been pursued in memory associations (Bruza et al., 2009; Nelson & McEvoy, 2007) and in decision making (Aerts et al., 2015; Basieva et al., 2019).

There has been no research exploring Bell's ideas for interacting agents. Our purpose is to develop a paradigm based on a PD variant, involving the interaction of a participant with a hypothetical counterpart. The payoff matrices can be set up in a way that optimal performance (relative to overall payoff) requires sensitivity to the counterpart's choices in some cases, but not others. Allowing participants to choose whether to communicate or not with their counterpart on every trial, we can examine participant's sensitivity to context and the capacity of different modeling approaches to capture behavior.

We propose two models for modeling choice behavior, based on the models widely employed in physics for Bell paradigms. The classical model (specifically, a local hidden variables one) is based on an assumption of perfect coordination between the interacting agents, but without communication of the questions each agent receives on any trial. It allows for no sensitivity to context. The quantum model is also based on an assumption of perfect coordination between the agents, but, additionally, it allows sensitivity to context up to a certain degree (quantified by Tsirelson’s bound, 1980). In physics, such quantum models are interesting, because they allow sensitivity to context, even though there is no obvious physical mechanism violating locality and free choice (and there is no signaling). In psychology, such a quantum model offers a particular hypothesis of the extent to which any communication between participants can translate to sensitivity to context.

Note, we could construct more elaborate classical models, in which the causal role of communication on the observed statistics is included, and such models could (in principle) be reconciled with the sort of paradigm we have outlined above. However, we think it is more interesting to explore a baseline classical model (perfect coordination, but no sensitivity to context) versus the standard quantum model (perfect coordination and some sensitivity to context), to inform our understanding of the extent to which participants could employ their information resource. We think it is surprising and interesting that, when $S>2$, as we shall see, a superficially reasonable classical model cannot offer a good description of behavior. Examining violations of Bell’s bound while allowing for interacting participants to break locality mimics attempts in physics to describe experimental statistics in Bell paradigms, by allowing violations of free choice and locality (Blasiak et al., 2021).

More generally, the use of QT in cognitive modeling follows an assumption that, in some cases, quantum principles offer better descriptions to human behavior (Busemeyer & Bruza, 2011; Pothos et al., 2013; Haven & Khrennikov, 2013). Quantum cognitive models have been explored for many kinds of cognitive processes, including decision making, categorization, similarity, perception, and memory. What is common amongst such diverse applications are a handful of characteristics which researchers have taken to be indicative of quantum-like processes. For example, sometimes behavior appears to be subject to interference effects, so that the law of total probability is violated – the PD games and analogous situations (Shafir & Tversky, 1992) are good examples. In other cases, when participants are asked to make a decision, it appears that the underlying mental state changes. Social psychologists have been aware of such processes for a long time (e.g., Schwarz, 2007). The added value from quantum models is that in QT there is a specific requirement for how the state ought to change as a result of measurements (in behavior, decisions) and various researchers have taken advantage of these processes to build cognitive models (e.g. Kvam et al., 2015; White et al., 2020). Of course, as outlined above, there have also been behavioral results indicative of sensitivity to context, for which the Bell framework and corresponding quantum models have been invoked to construct relevant theory (e.g., Aerts et al., 2016; Bruza et al., 2015). Quantum cognitive models have had good generative value, for example, in terms of anticipating biases from prior decisions (White et al., 2020) or a surprising constraint for question order effects (Wang et al., 2014).

As per our comments for Bell inequality violations above, in quantum cognitive models any quantum processes are epiphenomenal and are underwritten by an assumption of classical neurophysiology (Yearsley & Pothos, 2014). Moreover, there have been some compelling proposals of heuristic models, mimicking quantum models (e.g., Kellen, et al., 2018). So, why invoke the (unfamiliar) concepts of QT at all? There are two reasons. First, it appears that in some behavioural cases quantum models can offer particularly simple explanations. Such cases tend to be ones for which behaviour is sensitive to context (as in the present case) or there are conflicting biases for behaviour, which appear to interfere with each other. Second, different quantum models generally employ the same set of principles and so have been used to identify commonalities between findings which, up to that point, had been considered separate (e.g., Yearsley & Trueblood, 2018). So, even assuming that there is no 'real' quantum structure in the brain, and even if there are compelling mimicries between a specific quantum model and models based on other principles (as in Kellen et al., 2018), we think there is explanatory value in considering such models.

To summarise, we are aiming to provide a demonstration that a simple game involving two interacting agents can produce statistics which will challenge a baseline CPT model, but not (or less so) a quantum model. Conceptually, translating the Bell inequalities from a physics setting to a behavioural one has not been straightforward. Overall, we ran five experiments: the first three served a purpose of allowing us to refine and polish the methods to various degrees. The last two experiments (4 and 5) were more polished, in ways which will become apparent below.

2.4 Experiment 1

2.4.1 Method

2.4.1.1 Participants

Participants were recruited using Prolific Academic, restricting sampling to UK nationals. They were paid £2.25 for their involvement. Sample size was set a priori to 100; due to the format of online recruitment, the final sample was a little higher to 101. This sample had 50 females and 51 males. Participants were between 18 and 82 years old (M_{Age} = 35.75 years old, SD = 12.61). Participants also were asked about their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with the majority of the recruited participants reporting 5 (*n* = 97) and only a few others (*n* = 4) 4 or lower.

2.4.1.2 Materials

2.4.1.2.1 Prisoner's Dilemma

We employed a one-shot, PD variant, such that there were two possible questions for each player. Participants were told to imagine they were arrested with an associate and were both under suspicion for a minor crime in the Old Wild West. The sheriff of the town would interrogate them and their associate separately, asking one question to each. The sheriff would ask either: 1) "Did you know the victim?", or 2) "Were you at the scene of the crime?" The first question corresponds to *a*1 or *b*1 and the second to *a*2 or *b*2 (the participant's and their associate's questions are denoted by '*a*' and '*b*', respectively). Participants had two possible actions: to confess (equivalent to D in the standard PD paradigm; coded with a minus sign) or deny (equivalent to C; coded with a plus sign). Depending on the combination of questions, a different sentencing policy would apply. Participants were told that their sentencing policy would depend on their question and their response, as well as their associate's question and response (see Table 2.2). Participants were expected to favor decisions leading to lower sentences (fewer days spent in prison) for just themselves or for both themselves and their associate (Chater et al., 2008; Vlaev & Chater, 2006). We created 'Good' and 'Bad' payoff matrices, such that the sentencing would bias participants to deny or confess respectively. Note, participants were shown the payoff matrix just for themselves, but were told that their associate would receive the same payoff matrix.

Table 2.2 An example of a sentencing policy. In this case, a Deny decision by the participant will lead to a sentence of either 3 or 1 days in prison, depending on what their associate does.

	Associate Denies	Associate Confesses
You Deny	3	1
You Confess	24	23

There were eight unique trials which can be denoted as $a1b1$ good, $a1b1$ bad, etc. Each participant received all eight trials and was told to respond independently (e.g., each trial contained a different payoff matrix, and each associate had a different name across the trials). Table 2.2 shows an example of a good matrix in either $a1b1$, $a1b2$ or $a2b1$. For bad matrices, the payoff bias has been created with a bias towards confessing the crime. For good matrices, this bias is towards denying the crime. Note, the assumption that participants are responding independently may appear unrealistic. However, it is only marginally relevant to the present purpose, which was to collect data on choice behaviour ostensibly inconsistent with a simple classical model.

The participant's associate was hypothetical, and he/she was always assumed to behave as expected, e.g., in the case of trial $a1b1$ good, the associate would deny the crime in the b1 question. How would a participant know what the associate is likely to be doing? In most cases, there would be a choice associated with a lower sentence and so the participant would/ should guess that her associate would be selecting this option. This would be applicable for $a1b1$, $a1b2$, and $a2b1$ trials. For $a2b2$ trials, the payoffs would not uniquely identify an action as optimal. For these trials, the participant would receive a hint of what the associate is likely to be doing: participants were told that the sheriff does not know much about the crime, but he does know that exactly one between the participant and his/her associate, was at the scene of the crime. Participants therefore were cautioned that if the participant and his/her associate were to both confess or both deny for these trials, the sheriff would punish them with a high penalty. For example, for the $a2b2$ good trial, the sentencing matrix would be biasing towards anticorrelation between the participant and his/her associate, and the participant would receive an additional hint that the associate is 'likely' to deny the crime. So, sensitivity to context is built into the structure of the problem, in the simple sense that the participant's action needs to be informed by the associate's action when his/her question is $a2$, but not when it is $a1$. To clarify, given each payoff matrix, there is an ‘obvious’ response for what the hypothetical participant should be doing: we just assume that the hypothetical participant follows this action. The exception is the $a2b2$ case, where we offered an additional hint of what the hypothetical participant is doing.

On each trial, participants were allowed to choose whether to communicate (i.e., violate locality) or not (see Figure 2.1). They had the option to try to check, so as to discover the question that their counterpart was going to be asked. We discouraged participants from checking frequently by telling them that a check involved a risk of being caught and automatically receiving a high sentence. The first four trials in the experiment always attracted a penalty if a participant checked on his/her counterpart (these trials were fixed and different from the main experimental trials). Following these first four trials, without a noticeable break in the procedure, participants completed eight trials corresponding to each of the four combinations of questions in each of their Good/Bad instantiation. The recorded data concerned only these eight trials and participants never experienced the penalty for checking during these trials.

Figure 2.1. An example of the two sentencing policy matrices a participant would receive in the $a1b1$,$a1b2$, $a2b1$ conditions. In this case, knowledge of the counterpart’s question does not matter.

If Jenny is asked, “Did you know the victim?”, then you will receive:

	Jenny Denies	Jenny Confesses
You Deny	1	0
You Confess	20	19

Or, if Jenny is asked, “Were you at the scene of the crime?”, then you will receive:

	Jenny Denies	Jenny Confesses
You Deny	2	0
You Confess	20	17

However, for the *a*2*b*2 trials, whether participants checked on the associate or not would substantially change the sentencing policy. This was implemented by telling participants that the sheriff does not know much about the crime, but he does know that only one between the participant and his/her associate were at the scene of the crime. Participants therefore were cautioned that, for the *a*2*b*2 combination of questions (‘Were you are the scene of the crime?’), if the participant and his/her associate were to both confess or both deny, the sheriff would punish them with a high penalty (see the bottom of Figure 2.2). Note, in all *a*1*b*1, *a*1*b*2, *a*2*b*1 combinations of questions the sentencing policy would recommend a specific action (see the top of Figure 2.2). By contrast, for the *a*2*b*2 ones, this is not the case, since the bias for action from just the sentencing policy would be for the participant to anti-correlate with the associate. Therefore, for *a*2*b*2 trials, if a participant checked on his/her associate, there would be an additional hint regarding whether the associate would be likely to confess or deny the crime (in the form of the associate looking ‘nervous’ or ‘confident’).

Figure 2.2. An example of the two sentencing policy matrices a participant would receive in the *a*2*b*2 condition. In this case the payoff structure changes dramatically depending on which question the counterpart will receive.

If Lucy is asked, “Did you know the victim?”, then you will receive:

	Lucy Denies	Lucy Confesses
You Deny	15	8
You Confess	20	15

Or, if Lucy is asked, “Were you at the scene of the crime?“, then you will receive:

	Lucy Denies	Lucy Confesses
You Deny	100	2
You Confess	2	100

If a participant checked on their counterpart, they were told his/her question and were presented with the specific sentencing policy which would apply. For example, if a participant was contemplating the possibilities in Figure 2.2, and the trial was intended for the *a*1*b*1 combination of questions (question 1 for participant, question 1 for Jenny), and the participant decided to check on Jenny, he/she would be presented with the top matrix; that is, in this case, both the participant and his/her counterpart would be answering the question of whether they knew the victim. The participant then indicated his/her response for whether he/she wanted to deny or confess their crime. If a participant did not check on their counterpart, they were simply presented with both matrices and asked whether they wanted to deny or confess their crime.

2.4.1.2.2 Interpersonal Reactivity Index (IRI)

The purpose of this questionnaire was to explore whether the extent to which participants were sensitive to the contextual nature of the *a*2*b*2 trials depended on their degree of empathy. A relevant questionnaire for this purpose is the amended 14 item IRI (Davis, 1980), which asks participants to state how well a series of statements describes them (e.g., "I am often quite touched by things that I see happen" & "When I see someone being taken advantage of, I feel kind of protective toward them") on a five-point scale of *Does not describe me well* (0) to *Describes me very well* (4). The IRI has the following scales: Fantasy, Personal Distress, Perspective Taking and Empathic Concern. We did not compute the 'fantasy' and 'personal distress' subscales as they have been found to measure other constructs, rather than empathy (Baron-Cohen & Wheelwright, 2004; Lawrence et al., 2004; Spreng et al., 2009). Instead, we used the two subscales: Empathic Concern, and its usefulness in predicting co-operation or defection in the PD game (Batson & Ahmad, 2001; Batson & Moran, 1999; Cohen, 2010); and Perspective Taking, which has been shown to elicit distrust and selfish behaviour in mixed-motive contexts, increasing defection (Epley et al., 2006).

2.4.1.2.3 Cognitive Reflection Task (CRT)

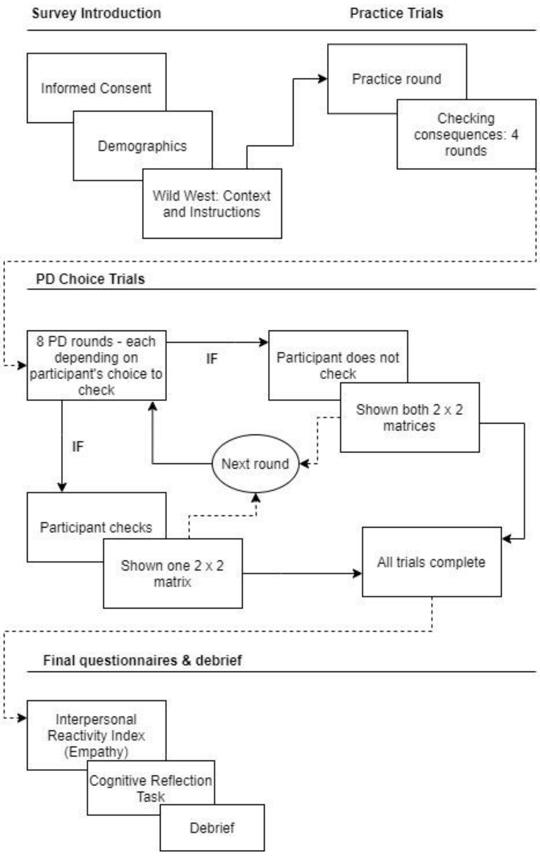
The CRT measures the tendency to override immediate incorrect answers and instead to elicit further reflections which lead to correct responses. The task contains three arithmetic problem-solving tasks, allowing participants to score between zero (low) and three (high). The CRT is a well-known measure of fast (system 1) versus slow cognition (system 2). Fast versus slow cognition can vary across individuals as a stable individual differences characteristic, but it can also vary for the same person, in terms of the effort he/she invests in a particular task (Pennycook et al., 2016). For example, Pennycook et al. (2016) highlight investment in non-intuitive tasks, such as strategies in chess, can negatively impact the propensity for one to think analytically in other contexts.

Given the format of the present paradigm, we can use the CRT to test for engagement and reflection with our PD tasks. However, the original CRT (Frederick, 2005) has been massively overused (Primi et al., 2016; Stieger & Reips, 2016). To reduce the likelihood that participants had encountered the original CRT in the past, we used three of the word problems presented in the appendices of Primi et al. (2016), who attempted to improve the psychometric properties of the CRT for adolescents and young adults. Participants read each of the questions and were asked to provide an answer in the text box.

2.4.1.3 Procedure

For an overview of the experimental procedure, see Figure 2.3. After giving informed consent, participants responded to some simple demographics questions and were then provided with some initial instructions which related to the context and particular format of the PD game we employed in the study.

Figure 2.3. Procedure for Experiments 1-3. Note, Experiments 2 and 3 did not include the IRI and CRT.


Survey Introduction
Practice Trials
Informed Consent
Demographics
Wild West: Context and Instructions
Practice round
Checking consequences: 4 rounds
PD Choice Trials
8 PD rounds - each depending on participant's choice to check
IF
Participant does not check
Shown both 2 x 2 matrices
IF
Next round
Participant checks
Shown one 2 x 2 matrix
All trials complete
Final questionnaires & debrief
Interpersonal Reactivity Index (Empathy)
Cognitive Reflection Task
Debrief

Following these instructions, participants responded to a few practice trials of the PD game, but with detailed additional instructions for each step of a trial. After these trials, participants were told that the main experiment would start. As noted before, to ensure that the potential consequence of checking on Bob was taken seriously, we included four consequence-checking rounds. In these rounds only, if participants checked on Bob they would receive an automatic 30 day prison sentence (participants were still told that there was a 'chance' that they would receive the harsher sentence if they checked on Bob). This was used to prevent participants from checking every round. Participants then completed eight PD trials (two rounds of questions $a1b1$, $a1b2$, $a2b1$, $a2b2$, such that for each case there was a version set for a response to confess and a version for a response to deny responses; trial order was randomised). Once completed, participants answered the 14-item IRI empathy questionnaire (Davis, 1980), before going through an updated 3-item CRT (Frederick, 2005; Primi et al., 2016). Participants were then provided with a debrief.

2.4.2 Results

Before conducting our main analyses (focussed on a quantity which measures super-correlation), some exploratory checks were implemented. First, we were interested in the extent to which participants checked on their counterpart when they were meant to, i.e., in the case of $a2b2$ trials. For this purpose, a chi-square test of independence was calculated comparing checks in the different question combinations (see Table 2.4). We first considered whether there was a difference in the overall frequencies of checks versus no checks, which was the case, χ^2 (1, *n* = 808) = 211.28, *p* = <.001. It is worth highlighting here that our chi-square analyses in this chapter violate the assumption of independence, since every participant offers a check or no check in each question combination. Due to this violation, the chi square analyses, mentioned here and elsewhere in this chapter, are entirely exploratory and should be interpreted with caution. To compensate for this shortcoming, we do provide other more precise analyses (such as one's sensitivity to checking, via Signal Detection Theory; Stanislaw & Todorov, 1999). Please bear in mind this limitation in subsequent uses of chi square tests to illustrate the basic pattern of results.

A more pertinent analysis concerns the frequency of checks across different combinations of questions, for example, were there more checks in $a1b1$ Good trials versus $a1b2$ Good trials? A priori, we expected more checks in $a2b2$ trials than for any other trials, given that it was only in these cases when the sentencing policy could vary substantially depending on the specific question that the participant's counterpart was asked. Note, we carried out these comparisons so that Good question combinations were compared only with other Good question combinations and analogously for the Bad ones; it does not seem pertinent to compare Good combinations with Bad ones. As expected, the frequencies of checks for the $a2b2$ trials were significantly higher than for all other trials, while the frequencies of checks for the rest of the trials did not differ significantly (see Table 2.5).

Second, we computed a d' sensitivity score, a standardised measurement using the ratio of the number of hits (checks on $a2b2$) relative to the number of false alarms (checks on $a1b1$, $a1b2$ and $a2b1$). The d' coefficient was calculated as

$$d' = \Phi^{-1}(H) - \Phi^{-1}(F)\text{..}\text{Equation (3)}$$

where H and F are hits and false alarms, respectively, and the Φ^{-1} function converts raw scores to z scores by fitting a normal distribution (0,1 mean and standard deviation) to scores from each participant and then inverting (Stanislaw & Todorov, 1999). Hits are considered instances of checking when the participants are meant to be checking (on *a*2*b*2 trials) and false alarms instances of checking when there would be no need for participants to check (on *a*1*b*1, *a*1*b*2 and a2b1 trials). Note, the mean d' score was above the neutral point (0), suggesting participants checked more frequently on *a*2*b*2 (hits) than *a*1*b*1, *a*1*b*2, *a*2*b*1 (false alarms; $M_{d'}$ = 1.54, SD = 1.53). Then, correlational analyses were conducted between d' scores and empathy subscales (empathic concern and perspective taking). No significant relationship was found between d' scores and either empathic concern (r = .06, p = .59) or perspective taking (r = -.06, p = .52) measures, suggesting that empathy is not a relevant characteristic for the decision of participants to check or not on their counterpart.

We next considered whether any variance in the d' variable could be explained by the empathy subscales in combination with the CRT scores (recall, the latter was employed as a proxy measure of engagement with the task). We ran a hierarchical linear multiple regression analysis with Model 1 including CRT and Model 2 including, in addition, the empathy subscales. Both Model 2, F (3, 97) = 7.6, p = .001, R^2 = .19, and Model 1, F (1, 99) = 20.48, p <.001, R^2 = .17, were individually significant. For information, we note that Model 2 explained only an additional 2% of variance in d', F (2, 97) = 1.13, p = .33, compared to Model 1, showing that taking into account (assumed) participant engagement with the task through the CRT does not alter the picture regarding the lack of relevance of the empathy subscales.

Lastly, we identified an error in the logic of the Qualtrics survey. To reiterate, the experiment was programmed so that Good and Bad matrices reflected the assumed action of their counterpart, that is, in Good matrices the participant's counterpart is always assumed to Deny the crime and in Bad matrices the counterpart is always assumed to Confess. As intended, checking the columns corresponding to Good sentencing policies, the overwhelming majority of participants chose to Deny (indicated as a +). Note that each question combination for a particular version, Good or Bad, allows only two possible actions for a participant, Deny or Confess, so that corresponding probabilities sum to 1. For example, in *a*1*b*1Good trials, the majority (96%) of participants denied their crime, whilst a small number (4%) confessed their crime. Analogously, for Bad trials most participants chose to Confess. The error in our survey was found in *a*1*b*1Bad trials (highlighted in Table 2.3), where most participants denied their crime despite our expectation that they should have confessed. These observed probabilities stemmed from participants initially viewing two matrices for which the optimal response was to Confess, but subsequently seeing matrices for these trials for which the optimal response was to Deny. This error was fixed in our replication in Experiment 2 and, moreover, does not impact on the analyses regarding when participants were most likely to check on their counterpart.

Table 2.3. Observed probabilities for all question combinations in Experiment 1.

	$a1b1$ Good	$a1b1$ Bad	$a1b2$ Good	$a1b2$ Bad	$a2b1$ Good	$a2b1$ Bad	$a2b2$ Good	$a2b2$ Bad
Prob(Participant +, Bob +)	.96		.92		.91		.17	
Prob(Participant -, Bob +)	.04		.08		.09		.83	
Prob(Participant +, Bob -)		.70		.11		.11		.71
Prob(Participant -, Bob -)		.30		.89		.89		.29

2.5 Experiment 2

2.5.1 Method

2.5.1.1 Participants

Participants were recruited using Prolific Academic and we restricted sampling to UK nationals only. They were paid £2.25 for their involvement. Sample size was set a priori to 100 participants, simply based on a consideration of consistency with the previous experiment. This sample had 51 females and 49 males. Participants were between 18 and 71 years old (M_{Age} = 34.51 years old, SD = 13.59). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with the majority of participants reporting 5 (n = 94) and only a few others (n = 6) reporting 4 or lower.

2.5.1.2 Materials & Procedure

Following the results of Experiment 1, Experiment 2's materials and procedure was identical to that of Experiment 1, except for the removal of the empathy measures (IRI). Some additional corrections were also made to the Qualtrics logic to ensure that participants were viewing the appropriate matrices for each trial.

2.5.2 Results

Once again, we were interested in the extent to which participants check on their counterpart when they were meant to, i.e., in the case of *a*2*b*2 trials. We first confirmed that there was a difference in the overall proportion of trials when participants checked versus not checked; this turned out to be the case, as assessed by a chi-square test of independence, χ^2 (1, *n* = 800) = 296.4, *p* = <.001 (see Table 2.4).

The key question is whether the proportion on checks was higher for *a*2*b*2 question combinations versus the other question combinations. This again proved to be the case, as shown by comparing proportions of checks for all question combination pairs (Table 2.5). However, there was a notable issue in Experiment 2. For the *a*2*b*2 Good matrix not checking would be expected to bias towards one decision and checking towards another. However, for the *a*2*b*2 Bad matrix the expected bias for decision was the same whether a participant checked or not checked. Whilst this is not necessarily problematic, this feature of Experiment 2 makes it harder to consider whether super-correlations are possible without checking. As such, we conduct Experiment 3 to rectify this issue.

2.6 Experiment 3

2.6.1 Method

2.6.1.1 Participants

Participants were recruited using Prolific Academic and we restricted sampling to UK nationals only. They were paid £2.25 for their involvement. Sample size was set a priori to 100 participants, consistent with our previous experiments. This sample had 50 females and 50 males. Participants were between 18 and 67 years old (M_{Age} = 35.1 years old, SD = 11.35). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with the majority of participants reporting 5 (*n* = 95) and only a few others (*n* = 5) reporting 4 or lower.

2.6.1.2 Materials & Procedure

Experiment 3's materials and procedure were identical to that of Experiment 2, except for a minor adjustment to one of the *a*2*b*2Bad matrices (as explained at the end of Experiment 2). With Experiment 3, we changed the *a*2*b*2 Bad matrix so that opposite decision biases were expected with and without checking (see Figure 2.4 below for illustration).

Figure 2.4. Examples of *a*2*b*2Bad matrices when optimal to confess (left) and optimal to deny (right).

Optimal to confess

If Jack is asked, "Did you know the victim?", then you will receive:

	Jack Denies	Jack Confesses
You Deny	20	20
You Confess	10	0

Optimal to deny

If Jack is asked, "Did you know the victim?", then you will receive:

	Jack Denies	Jack Confesses
You Deny	10	0
You Confess	20	10

2.6.2 Results

Once again, we were interested in the extent to which participants check on their counterpart when they were meant to, i.e., in the case of *a*2*b*2 trials. We first confirmed that there was a difference in the overall proportion of trials when participants checked versus not checked; this turned out to be the case, as assessed by a chi-square test of independence, χ^2 (1, n = 800) = 217.7, p = <.001 (see Table 2.4).

Table 2.4. Chi square tests for comparisons of rates of checking for all Good and Bad question combinations in Experiments 1-3.

Good Matrices	Exp. 1	Exp. 2	Exp. 3
*a*1*b*1 – *a*1*b*2	0.63	0.5	0.06
*a*1*b*1 – *a*2*b*1	2.26	0.1	1.84

$a1b1$ – $a2b2$	62.35**	83.56***	70.09***
$a1b2$ – $a2b1$	0.52	0.–8	1.23
$a1b2$ – $a2b2$	52.83**	80.74***	67.27***
$a2b1$ – $a2b2$	44.37**	75.33***	54.42***
Bad Matrices			
$a1b1$ – $a1b2$	0.20	0.4	0.82
$a1b1$ – $a2b1$	3.26	0.2	1.70
$a1b1$ – $a2b2$	72.36***	79.45***	57.55***
$a1b2$ - $a2b1$	1.87	1.3	0.17
$a1b2$ - $a2b2$	66.96***	71.22***	47.65***
$a2b1$ - $a2b2$	50.34***	85.33***	43.2***
Notes: *p < 0.05, **p < .01, ***p < .001. The n for the tests are 101 for Experiment 1 and 100 for Experiments 2 and 3.			

The key question is whether the proportion on checks was higher for $a2b2$ question combinations versus the rest. This again proved to be the case, as shown by comparing proportions of checks for all question combination pairs (Table 2.5).

Now consider the limitations of the current experiment setup. One impactful drawback from the current paradigm is that trials were presented in a 2 x 2 matrix regardless of whether participants decided to check or not. So why does this matter? The reason for a participant checking on their partner is to gather information about the sentence payoffs. If you receive a 2 x 2 matrix and you know you're receiving the same payoffs as your counterpart, there would be no logical reasoning for checking (especially if there is a chance of receiving a penalty for checking).

In the next experiment, we instead present an initial 2 x 1 matrix for the participant's payoffs. If the participant decides to check, then he/she would be told which question was assigned to the associate ($b1$ or $b2$) and the matrix would expand to show the payoffs for all combination of answers for the participant and associate (as seen in Tables 2.1 and 2.2). If the participant did not decide to check, then he/she would just be shown again the initial 2 x 1 matrix for just their payoffs. This addresses the limitation in Experiment 3, since it would encourage checking only in situations where there is ambiguity, such as the $a2b2$ trials.

Table 2.5. Frequencies of checking for each question combination in Experiments 1-3.

	Deny (Good)				Confess (Bad)				
	*a*1*b*1	*a*1*b*2	*a*2*b*1	*a*2*b*2	*a*1*b*1	*a*1*b*2	*a*2*b*1	*a*2*b*2	
Check	13	17	21	68	10	12	19	69	Exp. 1
No Check	88	84	80	33	91	89	82	32	
Check	11	12	14	75	10	13	8	72	Exp. 2
No Check	89	88	86	25	90	87	92	28	
Check	8	9	14	65	9	13	15	60	Exp. 3
No Check	92	91	86	35	91	87	85	40	

2.7 Experiment 4

2.7.1 Method

2.7.1.1 Participants

Participants were recruited using Prolific Academic and we restricted sampling to UK nationals. They were paid £2.25 for their involvement. Sample size was set a priori to 100 participants (50 males, 49 females and 1 participant who self-identified as ‘other’). Participants were between 18 and 62 years old (M_{Age} = 31.08 years old, SD = 11.70). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with the majority of participants reporting 5 (*n* = 97) and only a few others (*n* = 3) reporting 4 or lower.

2.7.1.2 Materials & Procedure

Experiment 4 was identical to that of Experiment 3, except for how the payoff matrices were presented to participants depending on whether they checked on their counterpart or not (see Figures 2.5 and 2.6). In this experiment, for a particular trial (e.g., a1b2) the payoffs in the 2 x 1 matrices were the approximate average of the payoffs in the 2 x 2 one. For example, looking at the checking matrices for Isabel (top left in Figure 2.6), we can work out that (29 + 21) ÷ 2 = 25. Note, averaged decimal payoffs were rounded up to the nearest whole number. But we did not create true averages across different trials. That is, for trial *a*1*b*2, there would be a 2 x 1 matrix which would be the average of payoffs in a corresponding 2 x 2 one. But the $a1$ payoff would not be an average from the $a1b1$ and $a1b2$ payoff matrices. These considerations are somewhat unimportant (in any case, they are addressed in Experiment 5), rather what matters is the bias for action, which was to Deny in all good matrices and Confess in the bad ones.

As per Figure 2.5 below, some initial instructions explained the format of the PD game. Participants then responded to a few practice trials, but with detailed additional instructions for each step of a trial. After these trials, participants were told that the main experiment would start. They first received the four consequence-checking rounds, and then the eight PD trials, after which the experiment concluded.

Figure 2.5. Procedure for Experiment 4.

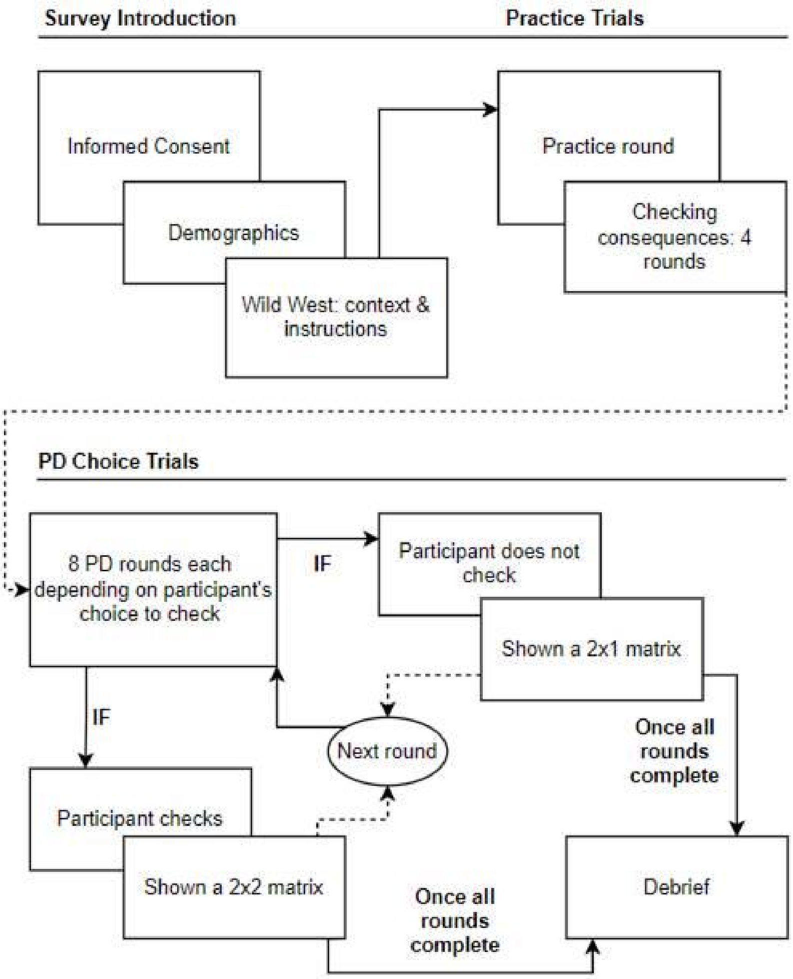


Figure 2.6. The Good and Bad matrices for $a2b2$ trials in Experiment 4.

*a*2*b*1 (bad)

Participant Does Not Check

You did not check on Isabel.

Isabel will be asked whether she knew the victim of the crime or whether she was at the scene of the crime. Since you don't know what Isabel's question will be, the following sentencing policy will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.

You Deny	27
You Confess	4

Were you at the scene of the crime?

Participant Checks

You checked on Isabel and found that she will be asked about whether she was at the scene of the crime. So, you know that the following policy for sentencing will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.

	Isabel Denies	Isabel Confesses
You Deny	29	21
You Confess	5	2

Were you at the scene of the crime?

*a*2*b*1 (good)

Participant Does Not Check

You did not check on Rick.

Rick will be asked whether he knew the victim of the crime or whether he was at the scene of the crime. Since you don't know what Rick's question will be, the following sentencing policy will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.

You Deny	3
You Confess	20

Were you at the scene of the crime?

Participant Checks

You checked on Rick and found that he will be asked about whether he knew the victim. So, you know that the following policy for sentencing will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.

	Rick Denies	Rick Confesses
You Deny	3	2
You Confess	20	19

Were you at the scene of the crime?

a2b2 (bad)

Participant Does Not Check

You did not check on Jack.

Jack will be asked whether he knew the victim of the crime or whether he was at the scene of the crime. Since you don't know what Jack's question will be, the following sentencing policy will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.

You Deny	52
You Confess	52

Were you at the scene of the crime?

Participant Checks

You checked on Jack and found that he will be asked about whether he knew the victim. So, you know that the following policy for sentencing will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.

	Jack Denies	Jack Confesses
You Deny	101	2
You Confess	2	101

a2b2 (good)

Participant Does Not Check

You did not check on Lucy.

Lucy will be asked whether she knew the victim of the crime or whether she was at the scene of the crime. Since you don't know what Lucy's question will be, the following sentencing policy will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.

You Deny	51
You Confess	51

Were you at the scene of the crime?

Participant Checks

You checked on Lucy and found that she will be asked about whether she knew the victim. So, you know that the following policy for sentencing will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.

	Lucy Denies	Lucy Confesses
You Deny	101	2
You Confess	2	101

Were you at the scene of the crime?	Were you at the scene of the crime?

2.7.2 Results

We observed a significant difference in the overall proportion of trials when participants checked versus not checked, χ^2 (1, n = 800) = 121.69, p = <.001 (Table 2.7). Additionally, participants were more likely to check with $a2b2$ trials than for other ones. Note, we carried out these comparisons so that Good question combinations were compared only with other Good question combinations and analogously for the Bad ones. Minimally, these results show that participants were sensitive to the context of their associate's questions, necessary to achieve higher performance.

We further show the choice probabilities to deny for all question combinations and separately for checking versus non-checking trials (Table 2.8). Consider the Deny/ Good/ Checking column. In this case, because the matrices are good, by design the participant's associate is meant to be denying; the participant should also recognize that it is better to deny. As expected, choice proportions reveal high probability for the participant to deny in pairs $a1b1$, $a1b2$, $a2b1$. For the last pair, $a2b2$, however, the participant and his/her associate are biased to anticorrelate and, given the associate will be denying, we observe a low proportion for deny choices, again as expected (.07). We observe the reverse pattern in the Deny/ Bad/ Checking column.

Note, when the participant is not checking, he/she ostensibly does not know which question the associate will be asked, and therefore there is no basis for the participant to distinguish between cases when he/she should correlate with the associate ($a1b1$, $a1b2$, $a2b1$) versus anticorrelate ($a2b2$). If this assumption were entirely correct, we should be observing identical deny proportions across all four question combinations, when not checking, but this is not the case (e.g., for the bad matrices, .65 is higher than the choice proportions for the other question combinations). As noted, the issue is that the reduced payoff matrix when not checking should be identical for $a1b1$ and $a1b2$ (and likewise for $a2b1$ and $a2b2$), but this was not the case (because reduced payoff matrices were constructed separately for each question combination). We address this issue in Experiment 5.

Despite this point, and similar points for Experiments 1-4, the results were still useful, as they allowed insight into the various factors which impact on the robustness of these experiments. While in principle we could employ the results from all experiments for modeling and for exploring the question of whether the particular classical versus quantum models we will propose are adequate, we will focus on the results from Experiments 4 and 5 (partly for brevity). Relatedly, the main empirical result across these four experiments is perhaps unsurprising: participants seek more information when existing information is inadequate for a decision. On one level, this is certainly true, since the task was designed to incorporate sensitivity to context in a particular way. On another level, our objective is less so to offer a surprising empirical finding, but to show that choice probabilities from this seemingly innocuous situation cannot be modeled by a classical model incorporating the (assumed) perfect coordination between the participant and her associate.

2.8 Experiment 5

Experiment 4 showed that participants recognized that there would be different biases for action depending on whether their associate's question was $b1$ or $b2$. In this experiment, we constructed payoff matrices so that the reduced matrix for e.g. $a1$ would be the collapsed matrix across the $a1b1$ and $a1b2$ possibilities (Figure 2.8).

Whereas previously there were only eight main trials (four question combinations in good and bad versions), for which participants were free to decide whether to check or not check, in this experiment we added eight trials when participants were forced to check and another eight trials in which participants were forced to not check (e.g., on some trials they were told that they had to check on their associate). Recall, to use Equation (1), we need probabilities for e.g. $Prob(++|a1b1)$, which is computed by considering the number of times the participant denies when given question $a1$ together with his/her counterpart denying when given question $b1$. Trivially,

$$Prob(++\,good\ |a1b1\ checking) = \frac{counts\ of\ deny\ a1b1\ good\ checking}{all\ counts\ of\ a1b1\ good\ checking}$$

With this approach, we can compute S values for the entire sample, but it is difficult to do so for individual participants, because e.g., a participant may have not checked in the case of the

$a1b1$ good trial. With the additional trials in this experiment, all relevant probabilities can be computed within participants, e.g.,

$$Prob(+ + good \mid a1b1\ checking) = \frac{counts\ of\ deny\ a1b1\ good\ checking\ for\ the\ participant}{all\ counts\ of\ a1b1\ good\ checking\ for\ the\ participant}$$

and so S values can be computed within participants (which enables us to conduct some statistical tests). For the example of this probability, $Prob(+ + good \mid a1b1\ checking)$, for a particular participant there would be a max of two relevant trials and a min of one trial, depending on whether the participant decided to check when he/she had the option to do so. We also introduced three questionnaires to this experiment: the Toronto Empathy Questionnaire (TEQ; Spreng et al., 2009), a Cognitive Uncertainty (CU) subscale from the Uncertainty Response Scale (Greco & Roger, 2001), and a variation of the CRT (Primi et al., 2016).

2.8.1 Method

2.8.1.1 Participants

Participants were recruited using Prolific Academic and we restricted sampling to UK nationals only. They were paid £4.50 for their involvement. Sample size was set a priori to 100 participants, and we recruited 101 participants, 50 males, 50 females and 1 participant who self-identified as ‘other’. Participants were between 18 and 78 years old (M_{Age} = 32.13 years old, SD = 12.54). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with the majority of participants reporting 5 (*n* = 95) and only a few others (*n* = 6) reporting 4 or lower. None of the participants for this experiment had taken part in Experiment 4.

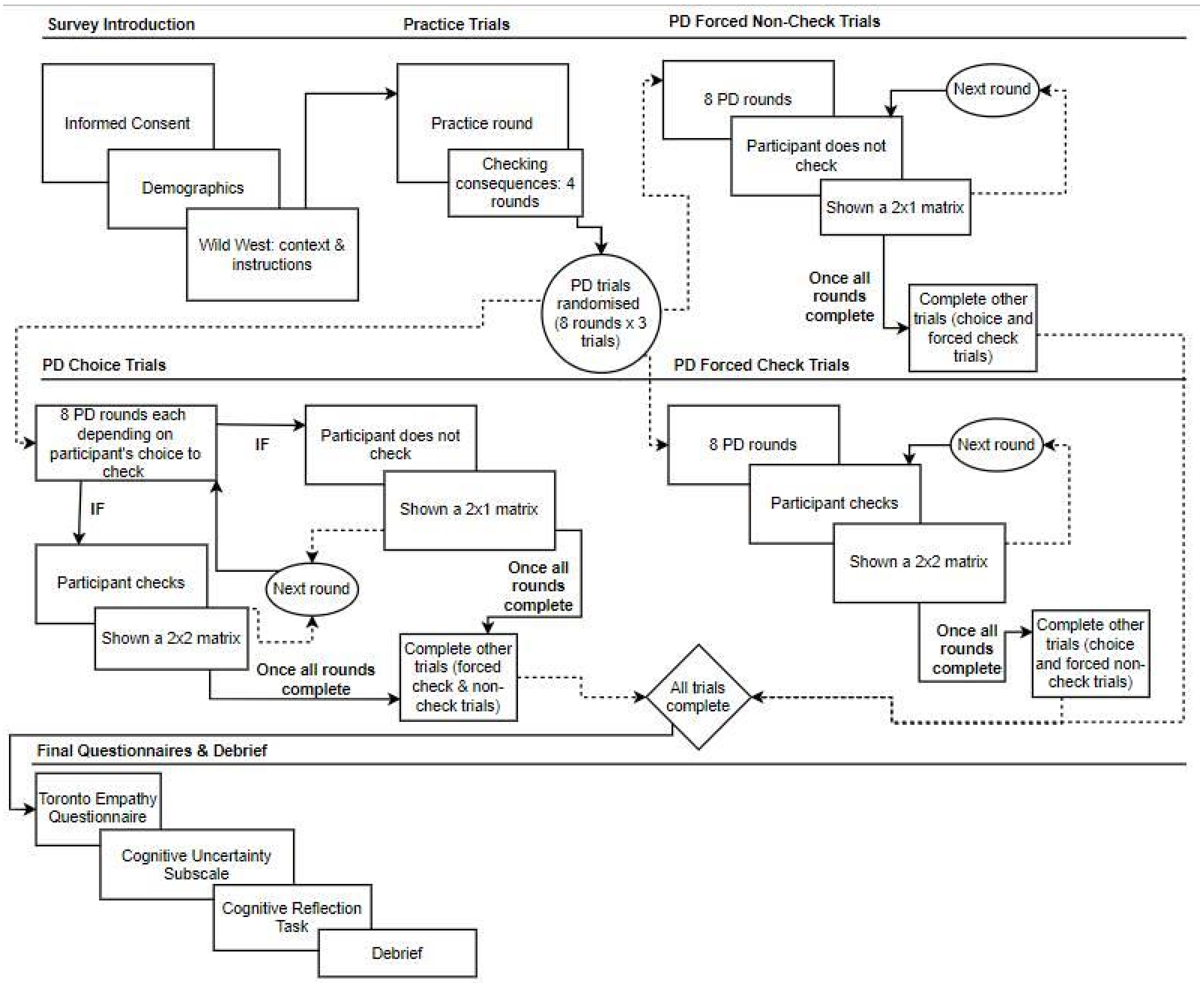
2.8.1.2 Materials & Procedure

In Experiment 5, the payoff matrices were set up so that if the participant did not check, the reduced 2 x 1 payoff matrix would be identical across the two possible question combinations, e.g. $a1b1$, $a1b2$ (Figure 2.8). Additionally, there were 24 trials in total: eight choice trials, where the participant can choose or not to check on their counterpart (as in Experiments 1-4), eight trials for which the participant is forced to check, and eight trials for which the participant is forced to not check. From a modelling viewpoint, this was advantageous as it enabled us to compare computations of S values in all three communication situations: 1) when participants do not check, 2) when participants do check, and 3) when all trials are considered.

We also included three questionnaires. First, we included the TEQ (Spreng et al., 2009), since the present task is one of guessing what a (hypothetical) associate is planning to do. We decided to measure empathy in this different way, because the trials for the corresponding experiment were more ambiguous, in terms of payoffs. Therefore, it is possible that decisions made in this paradigm may have been driven more so by the extent to which the participant is 'in tune' with their counterpart. In terms of the TEQ's psychometric properties, it has been generally praised for its construct validity, demonstrated via its associations with other measures of empathy, and its high internal consistency, ranging from ranging from $\alpha = .85$ to $\alpha = .87$, as well as its test-retest reliability at $r = .81$, $p < .001$ (Spreng et al., 2009). Moreover, it has also been validated across a number of countries and languages (Totan et al., 2012; Kim & Han, 2016; Kourmousi et al., 2017). So, and given our null results so far, it made sense to explore the usefulness of the TEQ in this context.

The questionnaire asks participants to rate 16 questions on a five-point scale, ranging from *Never (1), Rarely (2), Sometimes (3), Often (4) to Always (5)*. Items include, "Other people's misfortunes do not disturb me a great deal" and "It upsets me to see someone being treated disrespectfully". Second, we included the 17-item CU (Greco & Roger, 2001). The CU asks participants to state how well a series of statements describe them, including, "I like to plan ahead in detail rather than leaving things to chance" and "I like to know exactly what I'm going to do next" on a four-point scale of *Never (1), Sometimes (2), Often (3) and Always (4)*. This questionnaire assesses the possibility that checking behaviour is driven by uncertainty aversion. Finally, we employed the CRT to test for engagement and reflection with our PD tasks. However, the original CRT has been massively overused (Frederick, 2005; Primi et al., 2016). To reduce the likelihood that participants had encountered the original CRT in the past, we used three of the word problems presented in the appendices of (Primi et al., 2016). Participants read each of the questions and were asked to provide an answer in the text box. Note, Figure 2.7 details the full procedure of Experiment 5.

Figure 2.7. Procedure for Experiment 5.


Survey Introduction
Practice Trials
PD Forced Non-Check Trials
Informed Consent
Demographics
Wild West: context & instructions
Practice round
Checking consequences: 4 rounds
PD trials randomised (8 rounds x 3 trials)
8 PD rounds
Participant does not check
Shown a 2x1 matrix
Next round
Once all rounds complete
Complete other trials (choice and forced check trials)
PD Choice Trials
PD Forced Check Trials
8 PD rounds each depending on participant's choice to check
IF
Participant does not check
Shown a 2x1 matrix
IF
Participant checks
Shown a 2x2 matrix
Next round
Once all rounds complete
Once all rounds complete
Complete other trials (forced check & non-check trials)
All trials complete
8 PD rounds
Participant checks
Shown a 2x2 matrix
Next round
Once all rounds complete
Complete other trials (choice and forced non-check trials)
Final Questionnaires & Debrief
Toronto Empathy Questionnaire
Cognitive Uncertainty Subscale
Cognitive Reflection Task
Debrief

2.8.2 Results

As expected, when participants did not check, choice proportions were nearly identical across matched pairs of question combinations (e.g., *a*1*b*1 good and *a*1*b*2 good, Table 2.6). Once again, we were interested in the extent to which participants check on their counterpart when they were meant to, notably in the case of *a*2*b*1 and *a*2*b*2 trials. For this experiment, this analysis will only examine the trials when participants could decide whether to check or not. We first confirmed that there was a difference in the overall proportion of trials when participants checked versus not checked, χ^2 (1, *n* = 808) = 148.31, *p* = <.001 (Table 2.7). Moreover, participants were more likely to check with *a*2*b*1 and *a*2*b*2 trials than for other ones.

We next consider the individual differences measures. We computed d' (see Equation 3), empathy (TEQ), aversion (CU), engagement/reflection (CRT) scores and *S* for each participant, using Equation 1 (focused on the trials when participants could choose whether to check or not). Note, for d', due to the small number of trials per participants, we had a large number of probabilities of 0 or 1, which we corrected by adding 1 to the number of trials and 0.5 to the counts of hits and false alarms (Stanislaw & Todorov, 1999, p.144). Indeed, participants checked more so on *a*2*b*1 and *a*2*b*2 trials (hits) than they did in the other trials (false alarms). This is evident from the mean d' (*M* = .995, SD = 1.25) being above zero. All measures were then correlated with each other, without a multiple comparisons correction, as the intention is exploratory. There are two notable results. First, there was no relationship between individual participant *S* scores and d', *r*=-.135, p=.18. Second, there was a negative relationship between *S* and empathy, *r*=-.23, p<.05. Higher values of *S* imply higher sensitivity to context, which in this case means that a participant is better at recognizing when he/she should reverse decisions, based on what his/her counterpart is doing. One possible explanation for this result is that participants higher in empathy try to over-guess their counterpart's action, at the expense of considering the statistical properties of the game. There were no other significant results.

Table 2.6. Frequencies of checking for each question combination in Experiments 4 and 5.

	Deny (Good)				Confess (Bad)				
	*a*1*b*1	*a*1*b*2	*a*2*b*1	*a*2*b*2	*a*1*b*1	*a*1*b*2	*a*2*b*1	*a*2*b*2	
Check	26	13	23	69	25	30	33	63	Exp. 4
No Check	74	87	77	31	75	70	67	37	
Check	27	27	67	64	20	23	63	73	Exp. 5
No Check	74	74	34	37	81	78	38	28	

Table 2.7. Chi square tests for comparing rates of checking for all Good and Bad question combinations in Experiments 4 and 5.

Good Matrices	Exp. 4	Exp. 5
*a*1*b*1 - *a*1*b*2	5.38*	0
*a*1*b*1 - *a*2*b*1	0.24	31.84***
*a*1*b*1 - *a*2*b*2	37.07***	27.38***
*a*1*b*2 - *a*2*b*1	3.39	31.84***

*a*1*b*2 - *a*2*b*2	64.82***	27.38***
*a*2*b*1 - *a*2*b*2	42.59***	0.2
Bad Matrices		
*a*1*b*1 - *a*1*b*2	0.63	0.27
*a*1*b*1 - *a*2*b*1	1.55	37.81***
*a*1*b*1 - *a*2*b*2	29.3***	55.97***
*a*1*b*2 - *a*2*b*1	0.21	32.40***
*a*1*b*2 - *a*2*b*2	21.89***	49.63***
*a*2*b*1 - *a*2*b*2	18.03***	2.25

Notes: **p* < 0.05, ***p* < .01, ****p* < .001. The n for the tests are 100 for Experiment 4 and 101 for Experiment 5.

Figure 2.8. The good and bad matrices for the *a2b1* and *a2b2* trials in Experiment 5.

*a*2*b*1 (bad)

Participant Does Not Check

You did not check on Isabel.

Isabel will be asked whether she knew the victim of the crime or whether she was at the scene of the crime. Since you don't know what Isabel's question will be, the following sentencing policy will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.

You Deny	52
You Confess	52

Were you at the scene of the crime?

Participant Checks

*a*2*b*1 (good)

Participant Does Not Check

You did not check on Rick.

Rick will be asked whether he knew the victim of the crime or whether he was at the scene of the crime. Since you don't know what Rick's question will be, the following sentencing policy will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.

You Deny	51
You Confess	51

Were you at the scene of the crime?

Participant Checks

<table>
<tr>
<td>
You checked on Isabel and found that she will be asked about whether she was at the scene of the crime. So, you know that the following policy for sentencing will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.
<table>
<tr><th></th><th>Isabel Denies</th><th>Isabel Confesses</th></tr>
<tr><td>You Deny</td><td>52</td><td>52</td></tr>
<tr><td>You Confess</td><td>2</td><td>102</td></tr>
</table>
Were you at the scene of the crime?
</td>
<td>
You checked on Rick and found that he will be asked about whether he knew the victim. So, you know that the following policy for sentencing will apply. Please note, the numbers in the sentencing policies refer to the number of days you will serve in prison.
<table>
<tr><th></th><th>Rick Denies</th><th>Rick Confesses</th></tr>
<tr><td>You Deny</td><td>2</td><td>100</td></tr>
<tr><td>You Confess</td><td>51</td><td>51</td></tr>
</table>
Were you at the scene of the crime?
</td>
</tr>
<tr>
<td>
<i>a</i>2<i>b</i>2 (bad)
<b>Participant Does Not Check</b>
…
<table>
<tr><td>You Deny</td><td>52</td></tr>
<tr><td>You Confess</td><td>52</td></tr>
</table>
…
<b>Participant Checks</b>
…
<table>
<tr><th></th><th>Jack Denies</th><th>Jack Confesses</th></tr>
<tr><td>You Deny</td><td>101</td><td>3</td></tr>
<tr><td>You Confess</td><td>3</td><td>101</td></tr>
</table>
…
</td>
<td>
<i>a</i>2<i>b</i>2 (good)
<b>Participant Does Not Check</b>
…
<table>
<tr><td>You Deny</td><td>51</td></tr>
<tr><td>You Confess</td><td>51</td></tr>
</table>
…
<b>Participant Checks</b>
…
<table>
<tr><th></th><th>Lucy Denies</th><th>Lucy Confesses</th></tr>
<tr><td>You Deny</td><td>100</td><td>2</td></tr>
<tr><td>You Confess</td><td>2</td><td>100</td></tr>
</table>
…
</td>
</tr>
</table>

Table 2.8. Observed probabilities for all question combinations in Experiment 4 (*n*=100 for each cell), split by decision to deny or confess and whether participants checked or not.

		Checking		No Checking	
		Deny	**Confess**	**Deny**	**Confess**
Good	*a*1*b*1	.92	.08	.95	.05
	*a*1*b*2	.92	.08	.98	.02
	*a*2*b*1	.83	.17	.97	.03
	*a*2*b*2	.07	.93	.74	.26
Bad	*a*1*b*1	.04	.96	.15	.85
	*a*1*b*2	.07	.93	.11	.89
	*a*2*b*1	.12	.88	.1	.9
	*a*2*b*2	.83	.17	.65	.35

Table 2.9. Observed probabilities for all forced question combinations in Experiment 5 (*n*=101 for each cell*),* split by decision to deny or confess.

		Checking		No Checking	
		Deny	**Confess**	**Deny**	**Confess**
Good	*a*1*b*1	.94	.06	.93	.07
	*a*1*b*2	.95	.05	.94	.06
	*a*2*b*1	.66	.34	.69	.31
	*a*2*b*2	.10	.90	.67	.33
Bad	*a*1*b*1	.05	.95	.04	.96
	*a*1*b*2	.05	.95	.06	.94
	*a*2*b*1	.39	.61	.67	.33
	*a*2*b*2	.78	.22	.68	.32

2.9 Modeling

It appears that participants are sensitive to the context of their associate's decisions in the PD variants we employed, but does this sensitivity to context push choice statistics beyond the descriptive adequacy of classical models (of a certain kind) and, if yes, in what way? This is the key research question in the present work. The aim of the two models we will shortly present is to describe as closely as possible average choice statistics across trials. Note, the purpose of Experiments 1-3 was primarily to finetune the paradigm and so results from these experiments were not included in our modeling.

The data produced by Experiments 4 and 5 has the form of eight probabilities, corresponding to the decision of the participants to deny (plus) or confess (minus), when encountering the different PD payoff matrices (sentencing policies). The recorded probabilities always correspond to the participant deciding to plus. Therefore, for the $a1b1$ good matrix, the observed probability is recorded as $Prob(+\,+)$ and in the case of the $a1b1$ bad matrix, the observed probability is recorded as $Prob(+\,-)$. Of course, we further inferred $Prob(-\,+)$, $Prob(-\,-)$, etc.

We will present two models for the observed data, referred to as the classical hidden variables model (or just classical model) and the quantum model. It is more standard to formulate these models assuming a stochastic, rather than deterministic, associate. Accordingly, we combined choice statistics from the good and bad trials using e.g. $Prob(+\,+\,|a1b1) = Prob(+\,+\,|a1b1\ Good) \cdot Prob(Good|a1b1) + Prob(+\,+\,|a1b1\ Bad) \cdot Prob(Bad|a1b1) = Prob(+\,+\,|a1b1\ Good) \cdot Prob(Good|a1b1),$ where $Prob(Good|a1b1)$, $Prob(Bad|a1b1)$ refer to the probability of a good, bad game for a given choice of questions, respectively, and $Prob(+\,+\,|a1b1\ Bad) = 0$. Since in all cases we employed equal proportions of good, bad trials, for each choice of questions, then $Prob(Good|a1b1) = Prob(Bad|a1b1) = 0.5$.

2.9.1 Hidden variables classical model

According to this model, for each of the two agents, there is a hidden variable λ describing each sub-system, such that $\lambda_A = -\lambda_B$, with λ_A uniformly distributed over a 3D sphere. Note, this is an expression of perfect anti-correlation of the hidden variables corresponding to the agents, as opposed to perfect correlation, but this difference is immaterial. So, the main assumptions of the model are as follows. First, if the same questions are asked, the participant will always perfectly coordinate in the same way with the counterpart, that is, either always correlate or always anticorrelate; assuming always-correlation, if the participant denies, it is assumed the counterpart will deny as well etc. Second, there is a specific value for all question outcomes at all times. The implication of this more subtle assumption is that the participant should produce an outcome to her question, independently of which question is asked to her counterpart. In physics, this is the key realism assumption. Third, this model assumes locality and free choice. In the present experiments, we endow participants with a means of violating locality so, if they do this in a certain way, we expect the model to perform poorly. A final, minor, assumption is that the participant will generally recognize the optimal action in each trial (corresponding to a lower sentence), and she will always assume that her associate will also take the optimal action. This assumption is minor because of the way the payoff matrices were constructed, but if it is wrong, the model will just fail (both models will fail). In what follows, instead of a participant and her counterpart, we sometimes talk about two interacting agents, Alice and Bob.

The first agent is measured in two directions, $a1$, $a2$ and the second agent is measured in two different directions, $b1$, $b2$. In the present psychological context, 'directions' just correspond to the steer for action from each question, which is a function of the information in the payoff matrix and the agent's interpretation of this information (which will depend on his/her personality etc.). Non-trivial algebra shows that (e.g., Bell, 1964; note, the assumption concerning the existence of the hidden variable λ impacts on how these probabilities are derived):

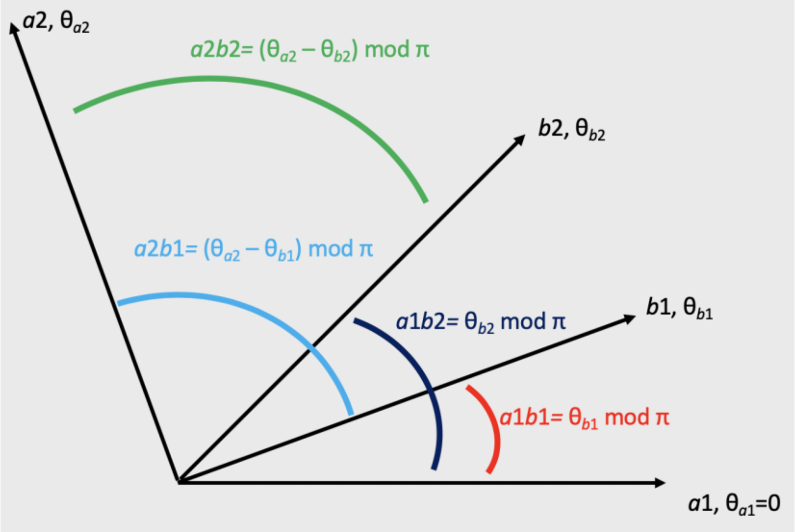
$Prob(++|a,b) = \frac{\theta}{2\pi} = Prob(--|a,b), Prob(+-|a,b) = \frac{1}{2} - \frac{\theta}{2\pi} =$
$Prob(-+|a,b)$...Equation (4)

The key parameter in Equation (4) is the angle θ, in radians, corresponding to the correlation between a measurement direction *a* for Alice and *b* for Bob. So, the joint probability for Alice and Bob to deny for question combination *ab* depends on the relation between how Alice perceives question *a* and Bob question *b*. Note that when $\theta = 0$, there is an equal chance for Alice and Bob to anticorrelate in one way (plus, minus) versus the opposite way (minus, plus), which is just an expression of the assumption $\lambda_A = -\lambda_B$, in the considered hidden variable model.

Since we have four pairs of measurement directions, $a1b1$, $a1b2$, $a2b1$, $a2b2$, then there are four angles as the parameters of this model. But these parameters are not independent. In the original physics set up they are actual measurement directions – psychologically, there is a corresponding assumption regarding the extent to which the two agents align or not in their consideration of questions. Suppose we have co-planar measurement directions, without much loss of generality. Then, the Figure 2.9 arrangement is a plausible representation of the four directions. Without loss of generality, we set θ_{a1}=0 and θ_{b1}, θ_{b2} and θ_{a2} as shown in Figure 2.9. Then, the four angles needed for the classical model are given as $a1b1$= θ_{b1} mod π, $a1b2$= θ_{b2} mod π, $a2b1$= (θ_{a2} – θ_{b1}) mod π, and $a2b2$= (θ_{a2} – θ_{b2}) mod π. The mod π function simply ensures that the angles for the four question pairs stay within the $0 < angle < \pi$ limit. It is defined as:

$$mod\ \pi(x) = \begin{cases} if\ x > 0, \begin{cases} if\ x - \pi < 0, x \\ if\ x - \pi > 0, 2\pi - x \end{cases} \\ if\ x < 0, \begin{cases} if\ x + \pi > 0, -x \\ if\ x + \pi < 0, 2\pi + x \end{cases} \end{cases} \quad \text{..Equation (5)}$$

Figure 2.9. The arrangement of the four measurement directions.


$a2$, θ_{a2}
$a2b2$= (θ_{a2} – θ_{b2}) mod π
$b2$, θ_{b2}
$a2b1$= (θ_{a2} – θ_{b1}) mod π
$a1b2$= θ_{b2} mod π
$b1$, θ_{b1}
$a1b1$= θ_{b1} mod π
$a1$, θ_{a1}=0

We next consider the S value given this classical model. $Prob(+ + |a1, b1\ )$ is the probability for both agents to +, when the questions are $a1$, $b1$, $Prob(+ - |a1, b1)$ the probability for Alice to + and Bob to – etc. Each expectation value is given by $\langle a\&b \rangle = \frac{2\theta}{\pi} - 1$, where θ is the angle between the measurement directions a, b. The overall result for the classical model is then:

$$S = \left| -2 + \frac{2}{\pi}[\theta_{a1b1} + \theta_{a1b2} + \theta_{a2b1} - \theta_{a2b2}] \right| \text{.....................................Equation (6).}$$

Note, we have mentioned that for this classical model S is bounded by 2. It can be shown that for $\theta_{a1b1} + \theta_{a1b2}$ the max is $2\pi - \theta$ and the min is θ, where θ is the angle between $b1$, $b2$, and for $\theta_{a2b1} - \theta_{a2b2}$ the max, min are θ and - θ. Together these results deliver the classical limits for S.

2.9.2 Quantum model

One of the most significant discoveries in the history of QT has been the capacity of the theory to break the classical $S \leq 2$ bound, seemingly without violating either locality or free choice. In the present paradigm, the situation is less philosophically challenging, since we endow the two agents with a communication capacity to break locality. Since the statistics produced by the quantum model are equivalent to classical ones, but with a degree of violation of locality (or free choice; Blasiak et al., 2021), the quantum model is a reasonable option for the present paradigm. The assumptions of the quantum model are equivalent to those of the classical one, but for two differences. First, instead of the Bayesian probability rules, we employ the probability rules from QT. Second, instead of a hidden variable capturing perfect coordination between the two agents, we have the quantum property of *entanglement* (see just below). However, this is not true (physical) quantum entanglement, but rather has epiphenomenal flavor (Yearsley & Pothos, 2014).

A column vector is denoted as $|x\rangle$, its conjugate transpose as $\langle x|$, and an inner product between two vectors as $\langle x|y\rangle$. Since we are concerned with two systems (agents), we need to employ tensor products to construct the joint state from the individual states, for example, $|x\rangle \otimes |y\rangle$ which can be written for brevity as $|xy\rangle$. We employ a qubit representation such that 0 means an intention for a '-' action (Confess) and 1 a '+' action (Deny). States are represented as $|\psi\rangle = a|x\rangle + b|y\rangle$. Measurements can change the state, so if on measuring ψ we obtain x the new state becomes $|\psi\rangle = |x\rangle$.

We start with state, $|\psi_+\rangle = \frac{|00\rangle - |11\rangle}{\sqrt{2}}$, where the tensor structure is so that the first index corresponds to Alice and the second to Bob (the subscript '+' in $|\psi_+\rangle$ simply indicates a 'correlation' state). So, $|00\rangle$ means that Alice is intending to minus and Bob to minus etc. Note, in physics, the state used is typically the singlet state, which is an anticorrelation state,

$|\psi_{-}\rangle = \frac{|01\rangle - |01\rangle}{\sqrt{2}}$. However, the predictions from $|\psi_{+}\rangle$ are essentially identical but for a fixed rotation of the measurement directions; so, for the purposes of model fitting, this issue is irrelevant (in a way analogous to that for the classical model). The state $|\psi_{+}\rangle$ is called entangled and is one of perfect coordination between the two agents, but now using the rules of QT. The predictions from the quantum model are then

$Prob(+ + |a, b; \psi_{-}) = \frac{1}{2}\sin^2\left(\frac{\theta}{2}\right) = Prob(- - |a, b; \psi_{-}), Prob(+ - |a, b; \psi_{-}) =$ $\frac{1}{2}\cos^2\left(\frac{\theta}{2}\right) = Prob(- + |a, b; \psi_{-})$...Equation (7)

As before, the crucial parameter is the angle θ for each measurement direction. The four angles are constrained as for the classical model (Figure 2.9), so that the quantum model also has three parameters.

We can consider the computation for the Bell bound from the quantum model. We have that the expectation values are given by $\langle a\&b\rangle = -\cos\theta$, where θ is the angle between the two measurement directions. Then,

$S = [-\cos\theta_{a1b1} - \cos\theta_{a1b2} - \cos\theta_{a2b1} - (-\cos\theta_{a2b2})]$ ....Equation (8)

It can be immediately seen that if we set the angle for $a1b1$, $a2b1$, $a1b2$ to $\frac{\pi}{4}$, with the arrangement as in Figure 2.9, $a2b2$ is $\frac{3\pi}{4}$. Then

$S = \left|-\cos\frac{\pi}{4} - \cos\frac{\pi}{4} - \cos\frac{\pi}{4} - (-\cos\frac{3\pi}{4})\right| = 2\sqrt{2} > 2$.................................Equation (9)

In fact, though not obvious from the present discussion, a quantum model cannot produce S values greater than $2\sqrt{2}$ and $2\sqrt{2}$ is called Tsirelson's bound (Tsirelson, 1980).

2.9.3 Overview of the two models

The question we are interested in is whether a model satisfying realism, locality, and free choice can model this data – this is the hidden variables classical model. The answer is not automatically no because, even though locality is violated, it is an empirical question whether participants recognize the need to employ non-local resources and use the available information efficiently. If participants do not employ the non-local information, then the results could still be described by a local model and S<2. That is, in this situation, the possibility of communication (checking) is clearly a necessary condition for participant data to violate Bell's bound, but it is not a sufficient one. A related question is whether any use of local information can be modelled by a quantum model (which is constrained by Tsirelson's bound) or not. If not, then participants' checking behavior and use of the corresponding information would be greater than what is allowed by QT.

Because in this case there is communication, it is likely there is signaling as well. If there is signaling, the bound of S=2 is clearly not a fundamental limitation on how a system behaves. However, there is still an empirical question on how people behave and we can ask the question (as above) of whether human behavior can be characterized by a local model (S<2), a nonlocal model constrained by Tsirelson's bound (the quantum model), or something else.

Table 2.10 shows the predictions from both models, where probabilities correspond to averaged data across multiple trials. This is easier to show by retaining the reference to the good, bad matrices, bearing in mind that in the fitted data we average probabilities across these two experimental situations, to better match the actual models.

Table 2.10. Correspondence between observed probabilities and predictions from the classical and quantum models.

Term	Observed probability	Classical prediction	Quantum prediction
$a1b1$ good	$Prob(+\,+)$	$Prob(++) = \frac{\theta}{2\pi}$	$Prob(+\,+) = \frac{1}{2}\sin^2\left(\frac{\theta}{2}\right)$
$a1b1$ bad	$Prob(+\,-)$ Bob will – in this case, but the probability we measure is for the participant to +.	$Prob(+-) = \frac{1}{2} - \frac{\theta}{2\pi}$	$Prob(+\,-) = \frac{1}{2}\cos^2\left(\frac{\theta}{2}\right)$
$a1b2$ good	$Prob(+\,+)$	$Prob(++) = \frac{\theta}{2\pi}$	$Prob(+\,+) = \frac{1}{2}\sin^2\left(\frac{\theta}{2}\right)$

$a1b1$ bad	$Prob(+-)$	$Prob(+-)=\frac{1}{2}-\frac{\theta}{2\pi}$	$Prob(+-)=\frac{1}{2}\cos^2\left(\frac{\theta}{2}\right)$
$a2b1$ good	$Prob(++)$	$Prob(++)=\frac{\theta}{2\pi}$	$Prob(++)=\frac{1}{2}\sin^2\left(\frac{\theta}{2}\right)$
$a2b1$ bad	$Prob(+-)$	$Prob(+-)=\frac{1}{2}-\frac{\theta}{2\pi}$	$Prob(+-)=\frac{1}{2}\cos^2\left(\frac{\theta}{2}\right)$
$a2b2$ good (Bob +'s)	$Prob(++)$ But recall this should now be low.	$Prob(++)=\frac{\theta}{2\pi}$	$Prob(++)=\frac{1}{2}\sin^2\left(\frac{\theta}{2}\right)$
$a2b2$ bad (Bob -'s)	$Prob(+-)$ This should be high.	$Prob(+-)=\frac{1}{2}-\frac{\theta}{2\pi}$	$Prob(+-)=\frac{1}{2}\cos^2\left(\frac{\theta}{2}\right)$

2.10 Model Fitting

Fits were assessed with Maximum Likelihood Estimation (MLE), using the G^2 expression for summary statistics in an experiment,

$$G^2 = 2N \sum_{trial\ types} \left(o_i \ln\frac{o_i}{e_i} + (1-o_i)\ln\frac{1-o_i}{1-e_i}\right)$$

$$= 2N \sum_{i,j=\{+,-\}} \left(Prob(ij, observed)\ln\frac{Prob(ij, observed)}{Prob(ij, model)}\right)$$

where N is the number of observations and o_i, e_i observed and expected probabilities for each trial type. Best fit for the models were identified through directed grid search with a step size for angle differences of 0.1; all parameters were taken to be uniformly distributed in a $[0,2\pi]$ range. For simplicity, since N was nearly identical for the two experiments, we ignored it in computing G^2.

2.11 Fit Results

Table 2.11 shows observed, classical predicted, and quantum predicted probabilities. Observe that for the $a1b1$, $a1b2$, and $a2b1$ pairs we recorded higher probabilities along the diagonals of the corresponding cells, but for the $a2b2$ pair, the opposite is true. This is the essential impression of supercorrelation and sensitivity to context: participants respond differently to question $a2$ depending on whether his/her counterpart received question $b1$ (correlation) versus $b2$ (anticorrelation).

We computed three S values, one for the observed choice probabilities, one for the predicted probabilities based on the classical model, and one for the predicted probabilities based on the quantum model. Note, for Experiment 5, empirical S was computed on the basis of the trials for which participants could freely choose whether to check or not on their associate. For Experiment 4, the empirical S, best fit classical S, and best fit quantum S were, respectively, 3, 2 ($G^2 = 0.46$), and 2.76 ($G^2 = 0.08$). For Experiment 5, the corresponding values were 2.46, 2 ($G^2 = 0.17$), and 2.65 ($G^2 = 0.09$). Bootstrapped 95% confidence intervals for the empirical S values were [2.73, 3.23] for Experiment 4 and [2.23, 2.71] for Experiment 5. The confidence intervals were computed by first calculating individual S values for each participant (only choice trials were used in this computation). Means were then calculated from each of the 1,000 bootstrap samples created (each bootstrapped sample was a random choice of N values from the original sample, with replacement, where N=number of values in the sample, i.e., the number of participants). Finally, the bootstrapped means were sorted and quantiles of .025 and .975 were utilized to indicate the 95% confidence intervals for each participant. In all cases, the empirical data show S>2, which demonstrates sensitivity to context and the impossibility of a four-way classical probability distribution to explain the data. The classical model resulted in worse fits than the quantum one, with the latter producing S values closer to the observed ones. Note that while the quantum model is able to capture a certain kind of sensitivity to context, of course it cannot describe any behavior (Tsirelson, 1980).

Using the forced checking and non-checking trials in Experiment 5, we computed S values for checking and non-checking trials for each participant. Note, in this case, it is only checking trials that should allow a violation of the $S \leq 2$ bound – therefore, for non-checking trials, it must be the case that $S \leq 2$. When participants were not checking on their associate, S for the good and bad trials respectively were 1.78 and 1.82; when checking, we observed 2.91 and 2.59 respectively. The difference in S between checking (averaged across good, bad matrices 2.75) and non-checking trials (averaged across good, bad matrices 1.80) was reliable, Z=-6.44, p<.001 (using the Wilcoxon Signed Rank Test, as the normality assumption would be suspect here).

Table 2.11. The observed and fitted results for Experiments 4 (left) and 5 (right).

	*b*1+	*b*1-	*b*2+	*b*2-		*b*1+	*b*1-	*b*2+	*b*2-
*a*1+	0.47	0.06	0.485	0.05	*a*1+	0.47	0.035	0.46	0.025
*a*1-	0.03	0.44	0.015	0.45	*a*1-	0.03	0.465	0.04	0.475
*a*2+	0.47	0.055	0.14	0.38	*a*2+	0.34	0.25	0.155	0.425
*a*2-	0.03	0.445	0.36	0.12	*a*2-	0.16	0.25	0.345	0.075
Empirical probabilities					Empirical probabilities				

	*b*1+	*b*1-	*b*2+	*b*2-		*b*1+	*b*1-	*b*2+	*b*2-
*a*1+	0.398	0.102	0.427	0.073	*a*1+	0.446	0.054	0.459	0.041
*a*1-	0.102	0.398	0.073	0.427	*a*1-	0.054	0.446	0.041	0.459
*a*2+	0.398	0.102	0.223	0.277	*a*2+	0.255	0.245	0.159	0.341
*a*2-	0.102	0.398	0.277	0.223	*a*2-	0.245	0.255	0.341	0.159
Probabilities predicted by the classical model					Probabilities predicted by the classical model				

	*b*1+	*b*1-	*b*2+	*b*2-		*b*1+	*b*1-	*b*2+	*b*2-
*a*1+	0.448	0.052	0.45	0.05	*a*1+	0.474	0.026	0.476	0.024
*a*1-	0.052	0.448	0.05	0.45	*a*1-	0.026	0.474	0.024	0.476
*a*2+	0.45	0.05	0.159	0.341	*a*2+	0.307	0.193	0.095	0.405
*a*2-	0.05	0.45	0.341	0.159	*a*2-	0.193	0.307	0.405	0.095
Probabilities predicted by the quantum model					Probabilities predicted by the quantum model				

2.12 Signaling

We finally, briefly consider the issue of signaling, for completeness. Signaling is an issue which is hugely significant in physics, but less relevant here. Briefly, signalling is a form of disturbance of one system on another. In physics, the aim is typically to identify violations of Bell without signalling (the quantum interaction between the two systems is also a form of disturbance, but more subtle than signalling). We can define a signaling quantity as:

$$I_S = \sum_{i=1,2} |\langle a_i\rangle_{b1} - \langle a_i\rangle_{b2}| + \sum_{j=1,2} |\langle b_i\rangle_{a1} - \langle b_i\rangle_{a2}|$$
$$= |\langle a_1\rangle_{b1} - \langle a_1\rangle_{b2}| + |\langle a_2\rangle_{b1} - \langle a_2\rangle_{b2}| + |\langle b_1\rangle_{a1} - \langle b_1\rangle_{a2}|$$
$$+ |\langle b_2\rangle_{a1} - \langle b_2\rangle_{a2}|$$

where the expectation values are defined as expected, for example, $\langle a_1\rangle_{b1} = (+1) \cdot \big(Prob(+ + |a1b1) + Prob(+ - |a1b1)\big) + (-1) \cdot \big(Prob(- + |a1b1) + Prob(— |a1b1)\big)$. Note, the max value for I_S is 8, when communication in both directions is considered (this is relevant in evaluating the size of the observed I_S values). We review a point which may lead to confusion: the probabilities in Tables 2.8 and 2.9 are not exactly the ones appearing in these expectation values. This is because, in Tables 2.8 and 2.9 we counted probabilities separately for the Good and Bad matrices, i.e., the probabilities in Tables 2.8 and 2.9 are e.g., $Prob(+ + |a1b1, Good)$. Therefore, as seen above too, we need to compute $Prob(+ + |a1b1) = Prob(+ + Good|a1b1) + Prob(+ + Bad|a1b1)$, but recall $Prob(+ + Bad|a1b1) = 0$. So, $Prob(+ + |a1b1) = Prob(+ + Good|a1b1) = Prob(+ + |a1b1, Good) \cdot Prob(Good|a1b1) = Prob(+ + |a1b1, Good)\frac{1}{2}$, because in the present design $Prob(Good|a1b1) = Prob(Bad|a1b1) = 1/2$ (meaning the probability of having a 'good' payoff matrix etc. ; the same applies for all question combinations). The probabilities $Prob(+ + |a1b1, Good)$ etc. are the ones in Tables 2.8, 2.9 and so in computing the expectation values for I_S, all probabilities from Tables 2.8, 2.9 need to be multiplied by a factor of ½ (the same applies to the calculations for the *S* values presented in Table 2.12).

We computed I_S separately for each experiment and for the checking versus no checking trials. For Experiment 4, for the checking and no checking trials we observed respectively $I_S = 0.08$ and $I_S = 0.33$. The corresponding values in Experiment 5 were $I_S = 0.08$ and $I_S = 0.04$. In Experiment 5, the results are as expected, since there is more signaling in the checking trials (ostensibly as a result of communication). In Experiment 4, even though for the no checking trials there was no communication, we still observed sizeable signaling. Signaling in Experiment 4 would be the result of the lack of balancing between the payoff matrices (as discussed in detail above). A consideration of signaling is clearly useful as a way to establish whether there might be unintended causal influences in the experimental statistics (as in Experiment 4). However, the non-zero I_S in Experiment 5, in the no checking trials indicates that signaling may be apparent even when there is no plausible corresponding mechanism, perhaps as a result of noise (Adenier & Khrennikov, 2017). This does recommend caution when employing signaling in such experiments, especially when the *N* is small (as would be the case in behavioral experiments).

The calculation of the signaling quantifiers I_S allows us to test for contextuality in the sense of Dzhafarov et al. (2016), which we do here for completeness. According to this work, contextuality is present whenever $|S| - I_S > 2$ (the *S* here refers to the maximum one between the four possible ways to compute it; here, we focused only on $S = |\langle a1\&b1\rangle + \langle a1\&b2\rangle + \langle a2\&b1\rangle - \langle a2\&b2\rangle|$, which is most relevant to our experimental design). In Table 2.12, we offer a complete record of relevant *S* values for the checking/ no checking quantifiers separately, for both experiments, as well as the quantities $|S| - I_S$, which are, as it happens, indicative of contextuality.

Table 2.12. Contextuality tests for Experiments 4 and 5.

	$\underline{S}$	$\underline{I_S}$	$|S| - I_S$
Exp 4 checking	3.2	0.08	3.12
Exp 4 no-checking	2.45	0.33	2.12
Exp 5 checking	2.74	0.18	2.56
Exp 5 no-checking	1.8	0.04	1.76

2.13 General discussion

Sensitivity to context is an important insight concerning the representation of information, whether in physics, data science, or psychology. Outside the physics of microscopic particles, it is assumed that there are no true quantum processes, and the study of sensitivity to context begs the question of the mechanism that supports it. In psychology, some pioneering work has been carried out so that both sets of questions, $\{a1, a2\}$ and $\{b1, b2\}$, would be answered by the same participant or in any case concern mental processes focused on the individual (e.g., Aerts et al., 2016; Bruza et al., 2009). Such approaches cannot be adapted to the interaction between separate agents, because, in general, without communication there is no possibility of breaking Bell's bound (or without rigging the choice of the questions asked to each agent).

For the first time, in this study we developed an approach enabling the application of the Bell framework in the interaction of two cognitive (and so macroscopic) agents. We considered putative locality violations as an information resource, that two interacting agents can employ at will (cf. Blasiak et al., 2021). We developed a simple empirical paradigm which embodied sensitivity to context in its structure, as a variant of a PD task (Shafir & Tversky, 1992). Empirical results showed that participants were sensitive to this context and the empirical *S* values exceeded Bell's bound. As noted, this is not surprising, given the structure of the payoff matrices we employed. The more surprising implication is that this sensitivity prevented fits by a simple classical model and so shows another way in which PD tasks and variants can produce results problematic for baseline expectation from CPT. 'Baseline' is a key qualification here since, as noted above, a classical model incorporating communication could be developed to account for the present results. Therefore, the present situation is not unlike most so-called paradoxes in probabilistic inference, for which a baseline classical probability approach appears erroneous, but it is always possible to offer accommodating elaborations (e.g., faced with a result such as *Prob*(X&Y)>*Prob*(X), one could write *Prob*(X&Y|A)>*Prob*(X|B)).

Theoretically, we fitted two closely matched models, a classical and a quantum one. The latter produced superior fits. This conclusion adds to the body of evidence that QT sometimes offers a good descriptive framework for behavior (Busemeyer & Bruza, 2011; Pothos et al., 2013). Elsewhere we have suggested that this is because QT looks like Bayesian inference, but in a local way (Pothos et al., 2021). That is, a set of questions for which it is impossible to have a complete joint probability distribution (e.g., because of resource limitations) is divided into subsets, such that within each subset – locally – we have Bayesian inference, but across subsets apparent classical errors arise. The idea that behavior is 'locally rational' has a precedent in psychology (Fernbach & Sloman, 2009; Lewandowsky et al., 2002).

Note, the immediate availability of locality violations to the participants makes it unlikely that any results showing *S*>2 would be due to 'correlations of the second kind', as discussed by S. Aerts and D. Aerts (S. Aerts, 2005; D. Aerts, 1990, 2014). In Experiment 5, when participants would check on the hypothetical counterpart we observed *S*>2 and when they would not *S*<2, showing that any apparent sensitivity to context was not brought about just by the measurements (decisions) themselves.

From the point of view of a physicist, the present results are interpreted as sensitivity to context, due to communication, regardless of whether this sensitivity to context is due to signaling or not. As noted, rather than considering signaling a nuisance influence, in this case we are interested in it, as a possible way in which Alice makes use of the information she has about Bob's questions.

There have been several challenges in realizing this project. First, the notion of applying the Bell framework to the interaction of cognitive agents superficially goes against the grain of Bell's work in physics. To address this problem, we had to formalize a notion of violations of locality or free choice, as information resources, which can be adopted versus not at will (formal work on this topic is reported in Blasiak et al., 2021), as well as consider the distinction between context sensitivity and contextuality (for the latter see Dzhafarov et al., 2016). Second, adapting the classical and quantum models developed for systems of microscopic particles in physics to behavioral data required careful consideration of the underlying assumptions of the models and how they could be matched to the behavioral situation. Third, the difference between contextuality and sensitivity to context and restrictive (or not) role of signaling in Bell-type paradigms are highly contentious issues. We think the approach we chose is justified, but equally we have offered additional analyses which we hope will allow researchers of differing opinions to still appreciate the results. Finally, reporting the research was challenging: the primary audience for this work is cognitive scientists, but we also hope to interest physicists and mathematicians familiar with Bell, who might be intrigued by applications outside physics. But, the mathematics is likely to be unfamiliar and challenging to cognitive scientists, while the details of the behavioral paradigm unfamiliar to physicists and mathematicians. Overall, one might say that interdisciplinary work of this kind, while conceptually exciting and potentially rewarding, is fraught with challenges – we can only hope that we have been at least partly successful in overcoming them.

The present analysis has practical potential. Consider two agents Alice and Bob, for whom it is in their interest to supercorrelate, but such that they are not meant to break locality and free choice, e.g., they are not meant to communicate. Alice and Bob might be an employee in a tech firm and a stockbroker considering investment opportunities in that firm, respectively. The present framework could be employed to determine whether Alice and Bob benefit from supercorrelation, either on the basis of violations of locality (which may reveal illegal insider trading) or free choice (which could correspond to Alice and Bob independently being sensitive to market conditions which determine the ‘questions’ each one of them has to respond to, at a given time). Clearly, the applicability of such an analysis depends largely on how the questions for each agent are specified and whether there is advantage in supercorrelation, which may not be very often.

In closing, we hope that the present work will further encourage researchers to employ the notion of contextuality and corresponding technical tools, in the study of the interaction between multiple agents.

Chapter Three: Constructive Influences

3.1 Introduction

In Chapter 2, we were able to identify some support for a quantum model, in the case of decision making between two interacting agents. A key characteristic of the quantum model is that a decision can change the system (the mental state of the participant). In this chapter, we develop this idea further.

Choice can be defined as a process of indicating a preference of one option over another. Indeed, a selected option is typically contingent on numerous factors, such as our previous experiences with the choices available (i.e., memory; Bordalo et al., 2017), the relative cost and benefits that each option confers (Jones & Hill, 1988; Fischhoff, 2015), and how the choices are presented to us (i.e., whether they are presented sequentially or concurrently; Gronlund et al., 2014; Trueblood & Dasari, 2017). Suppose we consider whether we like a film we just watched in the cinema, or whether we choose one dish over another in a restaurant, is this impression or decision already generated? Or do we play an active role in the constructive nature of deciding and evaluating?

There is a growing body of evidence to suggest that decision making is a constructive process, specifically that the process of making a choice can influence subsequent decision making (Schwarz, 2007; Sharot et al., 2010; White et al., 2014; White et al., 2016; White et al., 2019). One pertinent example comes from Botti et al. (2009), who examined highly undesirable, highly consequential decisions, such as deciding the medical treatment for a loved one in a vegetative state. Notably, they explored whether the opportunity for choice in the decision-making process can impact future outcomes of the families affected (i.e., a physician delegates decision making to the family, or the physician makes the decision for the family). Botti et al. found that those making tragic decisions generated more negative feelings than when having the same choices were made for them by their doctors, and families also perceived the tragic outcome to be a result of their own decision. Those who made the tragic decisions also elicited poorer coping abilities and were emotionally burdened with the sad choice they had to make. Put simply, the act of deciding (or not) generates thoughts and reflections (cognitive representations) of the experiences around us, which, in turn, impacts future decision making.

Such constructive influences are also probably implicated in question order effects. Notably, Moore (2002) found that American Vice President Gore would be considered less honest if the previous question concerned the honesty of President Clinton (and vice versa). Importantly, a judgement made on the honesty of Gore hinged on the context created by the previous question (regarding the honesty of Clinton).

Constructive influences are also likely involved when making a judgement conditional on pieces of evidence ordered in a certain way (Bergus et al., 1998; McKenzie et al., 2002; Trueblood & Busemeyer, 2011). For example, Bergus et al. (1998) examined whether the order of clinical information would affect diagnostic judgements of family doctors. The doctors were initially presented with a description of the patient (e.g., description of symptoms), and were then randomised to receive questionnaires that presented the same clinical information in one of two orders: the history and physical information first then lab results, or vice versa (note that the doctors were asked to provide a diagnosis at each information stage in the experiment). Although both groups of doctors were equally likely to successfully diagnose the patient from the information provided, Bergus et al. found that doctors were more likely to successfully diagnose a patient after receiving lab results first before receiving the history and physical information. Similar studies have also been conducted in jury decision making tasks. Researchers have found that the presentation order of the evidence can impact the verdict of a defendant (McKenzie et al., 2002; Trueblood & Busemeyer, 2011). Consider also the question-behaviour effect, according to which merely asking a question about intentions can influence subsequent actions on that behaviour (Sherman, 1980; Wilding et al., 2016). Clearly the order in which information is presented, and the questions asked, can impact on an outcome, in a range of situations including situations of substantial importance to everyday life, such as medical diagnoses, political polling, and jury decisions. But how do we explain the constructive nature of decision making?

One influential explanation relates to how the first question can activate thoughts, which subsequently affect consideration of the second question (Schwarz, 2007). This intuition was formalised by White et al. (2014). Their model predicted constructive effects for affective evaluations on advertisements, whereby expressing an opinion on a positively or negatively valenced stimulus would impact on subsequent evaluations of new stimuli. Specifically, they found that if an individual evaluates a positive image (e.g., a baby laughing) first and then a negative image (e.g., a burning building), the negative image will be rated more negatively than if it was evaluated by itself. Likewise, an evaluation of a negative image first and then a positive image will show more positive evaluations than if the positive image was rated by itself. White et al. labelled these constructive effects the EB.

We will now briefly introduce some of the notation relevant to the calculation of the EB. First, PN and NP refer to the valence order of the images shown to the participants (P = Positive; N = Negative). In the White et al example above, an image of a baby laughing followed by a burning building would refer to PN. The terms SSSR and SSDR refer to the ratings ‘second stimulus single rating’ and ‘second stimulus double rating’. Thus, to compute the EB, we note:

$$EB_i = (SSSR_{PNi} - SSDR_{PNi}) + (SSDR_{NPi} - SSSR_{NPi}) \text{ ....Equation (10)}$$

where i refers to the ith participant. The direction of the subtractions is based on the expectation that in the PN condition the intermediate rating leads to a more negative evaluation for the second stimuli and in the NP condition a more positive evaluation for the second stimulus. That is, when the EB is computed in this way, positive values are consistent with a certain prediction.

A few studies have now replicated White et al.’s (2014) demonstration with different stimuli and judgements. For example, White et al. (2016) observed an EB when asking participants to consider the trustworthiness of celebrities. In alignment with the observation that an evaluation for an advert would be stronger if followed by a previous evaluation, they found that when a more trustworthy celebrity was rated first, the trustworthiness of the second celebrity was lower than without the intermediate rating (and vice versa). The EB has also been found when asking employees to respond to questions on the state and/or strategy of their company (White et al., 2019).

But under what conditions does an individual exhibit the EB? One explanation could relate to whether an individual self-reflects on their own thoughts and behaviour; specifically, an awareness of how content makes one feel could drive the constructive influences reflected in the EB. For example, do some participants have more of a sense of the reasons of why/how they rate the first image? It is possible that what drives constructive effects in judgements is exactly the extent to which we are aware that we are expressing an opinion and the reasons for why we end up expressing a particular opinion. In this work, we use mindfulness and metacognition as measures of the extent to which we are aware of our own thoughts and behaviour. It is worth pointing out that the two notions are not entirely dissimilar. Hussain (2015, p.132) notes that: "[Although] the concept of meta-cognition and mindfulness seem different … meta-cognition and mindfulness share many commonalities and are conceptually related in many ways." As such, we will first provide a short overview of metacognition before we turn to mindfulness and the hypotheses for our experiments.

3.1.1 Metacognition

Humans are constantly monitoring their own thinking. From engaging in conversation to buying a gift for a friend to solving a complicated puzzle, we are constantly (voluntarily or involuntarily) reflecting on our own thinking. This phenomenon is known as metacognition and is a purely individual phenomenon, in which others play no direct role (Efklides, 2008). It also provides the obvious advantage that we can monitor and control our own cognition and behaviour (Dunlosky & Metcalfe, 2008). For instance, a student may consider the difficulty of an assignment (monitoring), and the amount of effort they may wish to put into it (control). Alternatively, a chess player may speculate about the experience of their opponent (monitoring) and implement appropriate strategies as needed (control). Indeed, these concepts depend on one another since there would be little point in expending significant effort in a task in a uniform way; and, equally, it would be problematic in recognising a difficult task and putting no effort in it. Note, metacognition has been further trichotomised into metacognitive skills, experience, and knowledge, but these sub-components are beyond the scope of this work (see Efklides, 2008). Instead, we are interested in how metacognition can provide the introspective insights an individual needs about the reasons why different decisions are made (see also *cognition of cognition*, Flavell, 1979; Livingston, 2003; Norman et al., 2019).

Now let us briefly consider how this might apply to White et al.'s (2014) task. In a given trial, a participant would view two adverts sequentially and rate them either one after the other (double rating) or after viewing them both (single rating). Importantly, they would monitor the stimuli to identify how positive or negative the stimuli are. It is here we expect that the thinking processes for the double and single conditions diverge. In the former case, participants would monitor the first advert and rate it before going on to monitor the second advert and then rating it (using the PN example explained earlier, this would equate to $Prob(burning\ building|baby\ laughing)$). In the latter case, participants would view both adverts but monitor only the second advert with a view to provide a rating (or in the PN example: $Prob(burning\ building)$). So why might we expect differences in how people rate the adverts between the single and double conditions? In terms of metacognition, differences can manifest in the participant's evaluations from the differing levels of monitoring and control between the stimuli. For example, a participant rating the stimuli above might rate the burning building as less negative after rating the advert with a baby laughing (carry over effects). Alternatively, the burning building may be rated more negatively in contrast to the baby laughing. To further explore this line of thinking, we will extend the work of White et al. (2014) to include metacognition as a predictor of the EB.

Now let us consider different ways to measure metacognition. Previous research has employed several different metacognition measures, including Judgements of Learning (JOL), Feeling of Knowing (FOK), and confidence ratings (Schwartz, 1994; Fleming & Dolan, 2012). There are also other measures, such as Ease of Learning (EOL) and Source Monitoring (SM), but we instead focus explicitly on JOL and FOK due to their popularity and use in the literature. The interested reader may consult Jemstedt et al. (2017) for EOL and Johnson et al. (1993) for SM. We will first examine JOLs and FOKs before outlining confidence ratings and why they were used in our experiments.

JOLs assess the likelihood that an individual can later recall learned material (Arbuckle & Cuddy, 1969; Rhodes, 2016). In a typical JOL experiment, participants are presented with word pairs, e.g., 'table-chair', and would be asked the likelihood they could recall 'chair' when later presented only with the word 'table' (Rhodes, 2016). Many researchers have utilised this paradigm and have demonstrated that JOL magnitude can be stronger when the words are related (e.g., knife-fork) rather than non-related (e.g., pen-dog; Castel et al., 2007); concrete (e.g., television) rather than abstract (e.g., acceptance; Tauber & Rhodes, 2012); and, recalled after a delay rather than immediately recalled (see Nelson & Dunlosky's (1991) *delayed JOL effect*; Rhodes & Tauber, 2011). Metamemory judgements (such as JOLs) are not only made in the encoding processes of memory, but they are also made at the time of retrieval via FOK (Nelson & Narens, 1994).

FOKs are predictions one makes on their ability to remember information, specifically whether information exists within their memory (Hart, 1965; MacLaverty & Hertzog, 2009; Fleming & Dolan, 2012). For instance, one can consider something they definitely know (e.g., one's own name); something they definitely do not know (e.g., Boris Johnson's phone number, at least for most people); and something they know an answer to but might be unable to retrieve the correct answer immediately (see *tip of the tongue experiences* in Brown, 1991). As such, previous research has studied FOK judgements using the Recall-Judge-Recognise (RJR) procedure (see Hart, 1965). In the RJR procedure, participants are presented with a series of questions and, for the questions they are unable to answer, they are asked to provide judgements of whether they could recognise the correct response from among wrong alternatives. Participants are then given a multiple-choice questionnaire with the same items presented in the initial test, and the answers are compared between the two response phases. Critically, the key question is whether FOKs are accurate when they occur.

Initially, researchers computed FOK judgements from memory tasks (as above and in Hart, 1965), but more recently FOKs have been captured by assessing Goodman-Kruskal gamma correlations (Nelson, 1984; Schnyer et al., 2004; El Saadawi et al., 2010), d prime (Higham et al., 2009), ROC analysis (Galvin et al., 2003) and meta-d prime (Maniscalco & Lau, 2014). Note, Goodman-Kruskal's gamma is a non-parametric measure of the strength and direction of association between two (or more) ordinal variables and is the most widely used measure in the metacognitive literature (Higham & Higham, 2019). Nevertheless, Fleming and Lau (2014) also observe that, "[these measures can all] be reduced to operations on [the] joint probability distribution: P(confidence, accuracy)."

We finally examine metacognitive confidence ratings. Metacognitive confidence refers to a judgement of certainty in relation to one's own performance on a task. Stankov et al. (2015) distinguishes confidence measurements into two types of assessments: 1) personality questionnaires assessing one's belief in their ability to complete different tasks, and 2) judgements of accuracy on a given task. The current experiment will focus on the latter and its relationship with metacognition. This topic has been extensively studied within forensic contexts such as eyewitness line-ups, and there have been findings of relatively stable links between confidence and accuracy when conducted under *pristine* testing conditions (for details on pristine testing conditions, see Loftus & Greenspan, 2017; Sporer et al., 1995; Wixted & Wells, 2017).

But what does good metacognition look like in relation to confidence? Carpenter et al. (2019, p.52) posit that an individual with good metacognition, "is aware of fluctuations in task performance, and appropriately modulates their confidence level (e.g., holding higher confidence when correct, and lower confidence when incorrect)." For example, an athlete may reflect upon their training performances before the Olympics. If the athlete was to regularly break their personal record in a 100-metre sprint, they may hold high confidence in their ability to perform in an upcoming race (and continue practicing their routines that brought them to that fitness level). Alternatively, if the athlete frequently recorded times lower than their personal best, it would likely warrant low confidence for their race (and changes to their training routine). On the other hand, poor metacognition would result from an incongruence between performance and confidence. An obvious benefit of utilising confidence ratings is that they are created during the task at hand, thus prompting immediate reflection on current performance. This enables us to observe confidence and accuracy as each trial progresses.

Accordingly, we measured metacognition using the average quadratic scoring rule (QSR, detailed in Experiment 1), adopted from Carpenter et al. (2019). The QSR enables us to compute a metric which measures how closely confidence ratings track accuracy. The way we employed the QSR involved two steps. First, we considered the absolute participant's (valence) rating for a stimulus and the average rating for the same stimulus from the corresponding pilot study. We converted these values into $\{0,1\}$, as a way to coarsen the data and reduce noise. Then, we used the QSR formula, producing a measure of discrepancy between subjective and 'objective' ratings. Without doubt, the validity of this approach to metacognition depends on noise in both the stimulus ratings and the confidence ratings. However, this is inevitable with any approach to metacognition along these lines; as long as the sources of noise are not systematic (e.g., they do not depend on the stimulus), then there should be no concern regarding the validity of the QSR. Metacognition could therefore be computed as a simple function of the extent to which confidence maps onto probability. For example, high confidence that an incorrect answer was actually correct would be indicative of poor metacognition; and, equally, low confidence that an incorrect answer was correct would indicate high metacognition.

In Carpenter et al.'s study, metacognitive ability was measured by providing participants with a perceptual task, measuring confidence, and looking at the relation between confidence and accuracy. They found evidence for a general capacity for metacognition with such tasks performed in different domains (e.g., perception versus memory) and that metacognition can be trained. Note, there were several differences between our experiments and the Carpenter et al. paper. First, Carpenter et al. ran a longitudinal study involving nearly 3,000 trials across the three phases of their study (pre-training = 432 trials; training sessions = 2160 trials; post-training = 432 trials). In our experiments, we employed 48 adverts (following White et al.'s paradigm) across two phases. The first phase involved rating how positive or negative the adverts were and the second phase concerned how confident the participant was with their rating. Second, meta d' was not computed in our experiments. This is because Carpenter et al. wanted to separately assess the effects of training on the metacognitive bias in their study, which was of no relevance to our experiments. Finally, we used a 0 to 100 slider scale (*0 being not confident and 100 being extremely confident*) instead of Carpenter et al.'s four-point confidence scale, on an assumption that this might enable participants to provide a more precise confidence response (DeSoto, 2014). Whether this difference makes an impact or not depends on whether naïve observers are able to provide granular ratings for their confidence. We assume they partly can, though this is an open empirical question.

3.1.2 Mindfulness

Mindfulness derives from eastern meditation techniques, notably Buddhism (Karunamuni & Weerasekera, 2019). It involves attending to one's current experiences and being situationally aware, without being overly reactive or overwhelmed by one's surroundings (Brown & Ryan, 2003). Mindfulness is often trained through meditative practices which allow for individuals to take note of their physical and psychological self in relation to their environment (e.g., yoga, walking meditation or mindful movements; Baer, 2003). Hölzel et al. (2011) note that mindfulness comprises of many 'synergistic' mechanisms which account for a process of enhanced self-regulation, leading to a range of positive psychological outcomes (Brown et al., 2007; Khoury et al., 2013). Examples of these mechanisms include an ability to decentre (or detach) from thoughts and emotions and re-perceive them as temporary rather than taking them as a fixed reality (Fisher et al., 2017); or, cognitive diffusion, whereby an individual reflects on their own thoughts without getting too caught up in them (Harris, 2009). It is unsurprising that mindfulness has been conceptualised as both a state (a temporary awareness of one's thoughts in the moment, e.g., Lau et al., 2006) and also as a trait (a disposition to be mindful in everyday life, e.g. Baer et al., 2006). Lindsay and Creswell (2017, p.49) extend these ideas, noting, "[Mindfulness is] a naturally occurring quality that varies across people (a disposition, or trait) and fluctuates across the day (a state of consciousness)." Indeed, Kiken et al. (2015) found that as an individual's state mindfulness increased, via meditation practice, trait mindfulness also increased over time, resulting in boosts to psychological health.

There have been numerous attempts at measuring mindfulness, including: The Mindful Attention Awareness Scale (MAAS; Brown & Ryan, 2003), Freiburg Mindfulness Inventory (FMI; Walach et al., 2006), and the Cognitive and Affective Mindfulness Scale (CAMS; Feldman et al., 2007). Another trait measure of mindfulness is the Five Facet Mindfulness Questionnaire (FFMQ), which was factorised from the aforementioned scales, and others (Baer et al., 2006).

We employed the FFMQ in our experiments for two reasons. First, it contains different facets, providing the advantage that we might discover differing associations between individual facets, the questionnaire composite measure, and the EB. Second, the instrument has been praised for its psychometric properties (Baer et al., 2006; 2008; Cebolla et al., 2012; Christopher et al., 2012; de Bruin et al., 2012).

The questionnaire assesses five subcomponents of mindfulness: Observe (O), Describe (D), Act with Awareness (AA), Non-Judge (NJ) and Non-Reactive (NR; Baer et al., 2006; 2008). The Observe subscale comprises of items relating to noticing internal and external experiences, such as sounds, emotions, thoughts, bodily sensations and smells. The Describe subscale refers to the ability to label one's experiences in words. The AA subscale involves attending to the present moment, rather than behaving automatically, while attention is focused elsewhere. The NJ subscale involves accepting and taking a non-evaluative approach toward thoughts and emotions (e.g., as "good" or "bad"). Lastly, the NR subscale refers to the ability to allow thoughts and emotions to come and go, without getting too involved or carried away by them.

Note, an important caveat of measuring state mindfulness is that the measurement occurs after (and not during) an experience. Interrupting the practice of mindfulness to take a measure of how mindful one is (e.g., how attentive or aware an individual is in each context) is inherently problematic to the measurement. It would be extremely challenging to be present and fully aware of your experience whilst responding to a question on your current level of mindfulness. Of course, this measurement issue is not unique to mindfulness studies and can be readily encountered within research on other psychological processes, such as recall in memory tasks.

We expect that those who are more mindful will exhibit reduced effects in the EB, which can be explained by two possible reasons. First, people who are more mindful may notice how the image makes them feel in the control condition, even in the absence of the question, eliminating the effect (Garland et al., 2015; Lindsay & Creswell, 2017). Second, people who are more mindful may be quicker to return to a baseline level of feeling so there would be less carryover effect (Creswell & Lindsay, 2014; Erisman & Roemer, 2010). This is not because they do not react to emotional stimuli, but rather because more mindful individuals are more able to move their attention quickly onto other things (i.e., they may be less likely to get 'stuck' on a thought or feeling). Indeed, if the latter point holds true, the timing of stimulus presentation may be important in revealing differences between those who exhibit the EB and those who do not (but note, as it turns out, we did not explore this angle).

3.1.3 Summary

Could constructive influences stem from an insight into the reasons for one's actions? This chapter aims to address this question. We will first attempt to replicate the EB and then we will explore whether different levels of self-reflection (in terms of metacognition and or mindfulness) impact an individual's propensity for constructive influences, as measured by the EB. Due to the exploratory nature of this study (e.g., differing relationships within facets of mindfulness), our expectations of how these variables will exactly interact are ambivalent. A positive relationship between Observe and the EB might suggest that attention to present moment experiences increase awareness of the stimuli, evoking a stronger emotional reaction between ratings. Viewing feelings as separate from oneself (decentering) might also reduce the EB. Similarly, perhaps those who score higher on overall mindfulness may be quicker to return to an emotional baseline, so there would be less carryover when rating the second advert, in turn, reducing the EB. But note there is a converse possibility, too, since people who are more mindful may more accurately observe the first stimulus and their own rating and show a greater effect for this reason. In terms of metacognition, our expectations are slightly more honed. Of course, White et al. found significant differences between single and double ratings, specifically noting that when participants provided an intermediate rating, they were more likely to exhibit a constructive effect. Here, we attempt to show that these significant differences can be related to a participant's metacognitive awareness (i.e., monitoring of the first rating impacting the second rating versus the second rating alone). We reserve more careful analysis for after the empirical results.

3.2 Pilot

We conducted a pilot study for two reasons. First, we wanted to check the valence of the adverts to be used in our experiments. We utilised the same database of images as White et al. (2014; via GAPED, Dan-Glauser & Scherer, 2011), and introduced some new advertisements intended to have neutral valence. It was important for us to ensure the valence of all adverts mapped onto their appropriate affect in this pilot study. Second, the affect ratings (mean score ±1SD) would also provide a useful benchmark for future experiments by helping us determine whether participants have accurately rated an advert's affect or not (see Experiment 1, metacognition, for more details).

3.2.1 Method

3.2.1.1 Participants

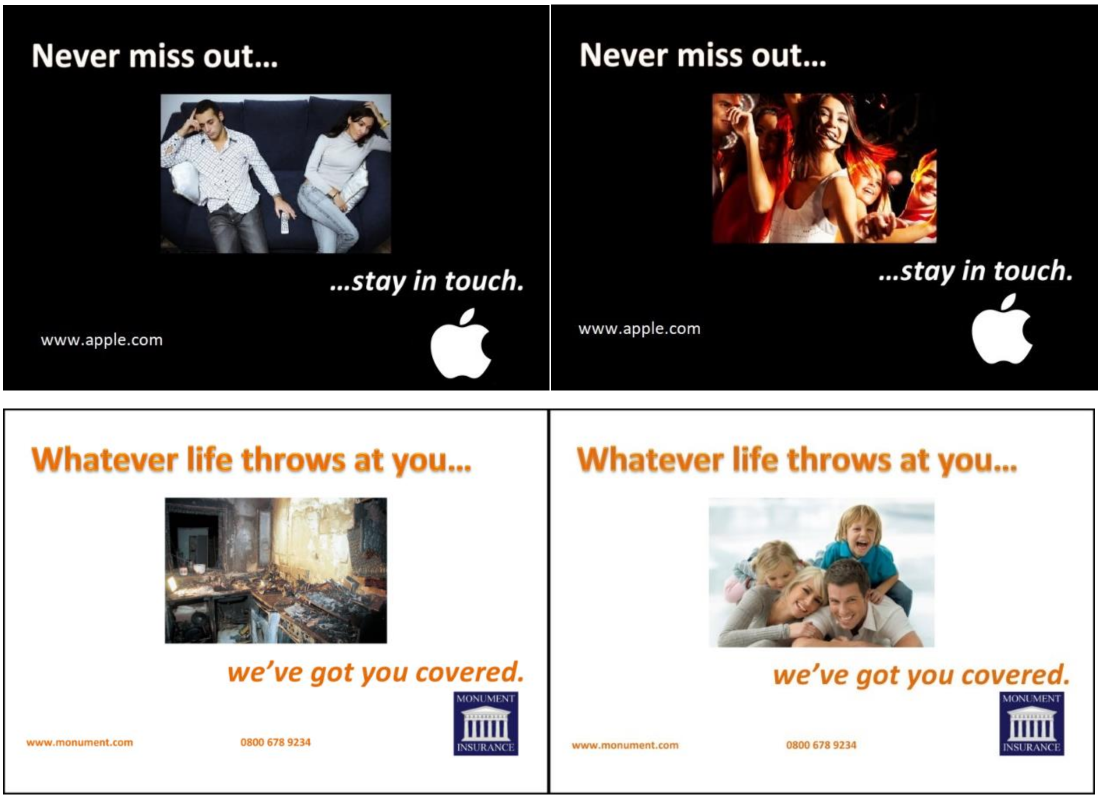
Two hundred participants were recruited using Prolific Academic, restricting sampling to UK nationals. They were paid £1.50 for their participation. Sample size was set a priori to 200; however, the final sample contained fewer participants due to outliers (*n* = 9, whose valence measures were ≥ 3 standard deviations). The final sample contained 95 males, 94 females, and 2 participants who indicated 'other' (*n* = 191). Participants were between 18 and 68 years old (M_{Age} = 33.98 years old, SD = 12.21). Participants were also asked about their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with many of the recruited participants reporting 5 (*n* = 180).

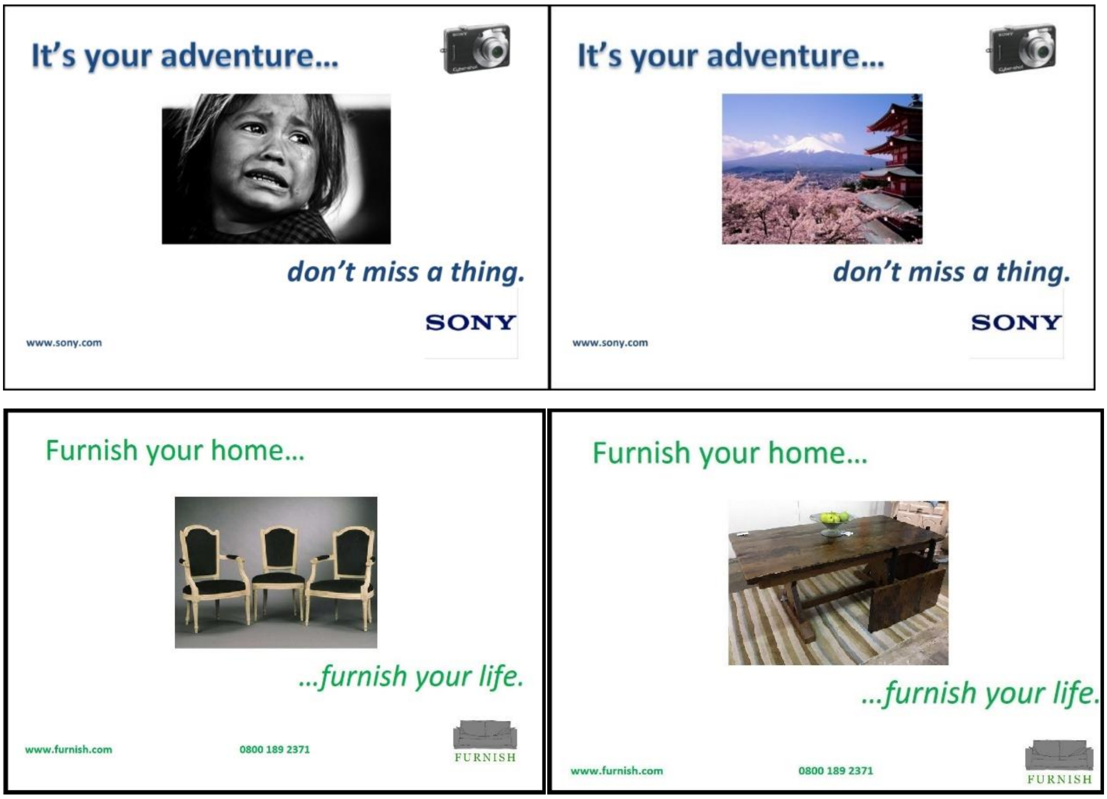
3.2.1.2 Materials

3.2.1.2.1 Adverts

Sixty adverts were selected for valence-checking, most of which (36) were from White et al. (2014) and the remainder (24) were new for this pilot. These images corresponded to four themes (insurance, mobile phones, cameras and furniture; see Figure 3.1 below), created with messages that fit with positive, negative and neutral valence. Images from White et al (2014) retained the same advert design (an exception to this was the mobile phone adverts; we changed the Blackberry adverts to Apple to make them more familiar to participants). The stimulus category that was new relative to White et al. (2014) was furniture and this category contained neutral fillers only.

Figure 3.1. Adverts shown in the Pilot study. The left side shows negative stimuli and the right side shows the positive stimuli, with the exception of the final row of neutral stimuli. The themes of the adverts are shown in the following order: mobile phones (Apple), insurance, cameras and furniture. Note, the only differences between the Apple and Blackberry mobile phone adverts (used in White et al.'s original experiment) is the website name and logo.





3.2.1.3 Procedure

After giving informed consent, participants responded to some simple demographics questions and were then provided with some initial instructions regarding the task. Participants were told they would view a series of adverts and would be asked, "How does this advert make you feel?", responding on a nine-point scale, from *1: Very Unhappy* to *9: Very Happy*. Each trial consisted of an advert, presented for five seconds to ensure participants had ample opportunity to view the advert, before a request for a rating appeared. The presentation order of the adverts was randomised. Once all trials were completed, participants were provided with a debrief.

3.2.2 Results

We conducted a series of paired samples t-tests to examine differences in the valence between the positive, neutral and negative images. First, images categorised as negative (M = 3.49, SD = 0.91) were rated significantly lower than images categorised as positive (M = 5.99, SD = 0.97; *t*(190) = -24.97, *p* < .001, *d* = 2.66). Second, images categorised as neutral (M = 4.49, SD = 0.72) were rated significantly higher than images categorised as negative (*t*(190) = -14.79, *p* < .001, *d* = 1.21). Lastly, images categorised as positive were rated significantly higher than images categorised as neutral (*t*(190) = 19.15, *p* < .001, *d* = 1.76). These results taken together indicate that the adverts map onto their intended valence (see Table 3.1), replicating the pilot study conducted by White and colleagues (2014).

Table 3.1 Valence for all adverts in the pilot study.

	Negative	Neutral	Positive
Mean	3.49	4.49	5.99
SD	0.69	0.88	0.5
Mean - 1SD	2.81	3.61	5.49
Mean + 1SD	4.18	5.37	6.48
# adverts	18	24	18

We next focused on the removal of six adverts from each of the categories (eighteen in total) with a valence beyond ± 1 SD from the mean of the stimuli. Since adverts were presented in pairs, some images were removed to ensure there were an even number of images. High scoring adverts were prioritised for removal for negative adverts, high or low scoring for neutral images and low scoring adverts were prioritised for positive images (recall, participants were asked how the adverts made them feel and to score the degree of positive or negative valence). To reduce divergence from the expected valence, eighteen adverts were removed, leaving 42 adverts available for subsequent studies. The mean affect ratings of the images from this pilot sample (±1SD) further enabled us to calculate a minimum and maximum range of affect for each image for future experiments, allowing us to determine whether participants have accurately rated an advert's affect or not (see Experiments 1, 2, metacognition and Table 3.2 below).

Table 3.2 Valence for remaining adverts after removing 18 adverts of more ambiguous valence in the pilot study. Note, these adverts were rated on a nine-point scale (1 = very unhappy; 9 = very happy).

	Negative	Neutral	Positive
Mean	3.13	4.22	6.22
SD	0.49	0.6	0.38
Mean - 1SD	2.64	3.62	5.84

Mean + 1SD	3.61	4.83	6.6
# adverts	12	18	12

3.3 Experiment 1

This experiment consisted of two parts. The first part consisted of a (nearly straight) replication of White et al. (2014), aiming to reveal the same constructive effect on judgement as in the original study. Second, we implemented measures of metacognition and mindfulness as possible predictors of whether an individual's propensity for constructive influences is moderated by their ability to self-reflect.

3.3.1 Method

3.3.1.1 Participants

Two hundred participants were recruited using Prolific Academic, restricting sampling to UK nationals. They were paid £3 for their participation. Sample size was set a priori to 200. However, due to the format of online recruitment, the final sample was a little higher at 201. The sample contained 101 males and 100 females. Participants were between 20 and 70 years old (M_{Age} = 39.97 years old, SD = 12.56; two participants did not disclose their age). Participants were also asked about their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with the majority of the recruited participants reporting 5 (*n* = 191) and the others (*n* = 10) reporting 4 or lower. Finally, we recruited equally for meditators and non-meditators using Prolific's pre-screening system. However, only 68 participants reported that they meditate in response to questions within our survey, $M_{\text{MeditationFrequency}}$ = 4.07 times per week, SD =3.29; $M_{\text{MeditationDuration}}$ = 22.79 minutes per session, SD = 18.8; the majority (~1SD) of meditators reported having between 9 months and 10 years' experience (see Table 3.3).

Table 3.3 Frequency of meditation experience in Experiment 1 (*n* = 68; *M* = 6.94, SD = 2.75).

Item Number	Meditation Experience	Frequency	Valid Percent	Cumulative Percent

1	1-2 months	4	5.9	5.9
2	3-4 months	3	4.4	10.3
3	5-6 months	6	8.8	19.1
4	7-8 months	2	2.9	22.1
5	9-10 months	1	1.5	23.5
6	11-12 months	1	1.5	25
7	1-2 years	16	23.5	48.5
8	3-4 years	15	22.1	70.6
9	5-10 years	11	16.2	86.8
10	11-20 years	5	7.4	94.1
11	21+ years	4	5.9	100
Notes: Mean and standard deviation reflect item number.				

3.3.1.2 Materials

3.3.1.2.1 Adverts

Forty-two adverts were employed. There were 12 positive, 12 negative and 18 neutral images, each selected on the basis of valence from the pilot study. To replicate White et al.'s (2014) design, all images were used in both single and double rating conditions. Eight adverts in the PN condition included two positive insurance, two negative insurance, two positive smartphone and two negative smartphone, and likewise for the NP condition. Positive and negative fillers were also used to create PP and NN trials (4 trials, involving 8 adverts, with 4 positive camera and 4 negative camera adverts). The images were randomly presented with 18 additional neutral adverts.

3.3.1.2.2 Five Facet Mindfulness Questionnaire

The 39-item FFMQ (Baer et al., 2006; 2008) requires participants to rate how accurate a series of statements reflects their experience, related to mindfulness, over the past month (e.g., "When I do things, my mind wanders off and I'm easily distracted." & "I don't pay attention to what I'm doing because I'm daydreaming, worrying, or otherwise distracted."). Responses were provided on a 5-point scale ranging from *1 (Never or very rarely true)* to *5*

(Very often or always true). The questionnaire assesses five facets of mindfulness, including: Observe, Describe, AA, NR and NJ (for a detailed reminder of each facet, see the Introduction).

3.3.1.2.3 Metacognition

To reiterate, metacognition refers to the ability to monitor and reflect on one's cognition; specifically, for the present study, we focus on cognitive performance. An individual with good metacognition will experience higher confidence when correct, and lower confidence when incorrect, in some task. In the current study, we employed the same 42 adverts from the White et al. (2014) task and asked participants to respond to a series of metacognitive judgements. First, participants viewed each advert individually and were asked to rate how positive/negative each advert was. Second, they were then asked to provide confidence ratings for each advert. To compute a metacognition score for each participant, we used the QSR, which measures how closely confidence ratings track accuracy (Carpenter et al., 2019):

$$QSR_i = 1 - (accuracy_i - p(correct)_i)^2$$

$$QSR = \sum_{i=1}^{all\ trials} 1 - (accuracy_i - p(correct)_i)^2 \text{................................Equation (11)}$$

The index *i* tracks the trials in the Metacognition Task. Accuracy refers to whether on the i^{th} trial the participant was correct or not and is a binary 0,1 variable. For each of the positive, negative and neutral advert types, a minimum/maximum valence range was created from the pilot adverts based on the means and standard deviations of the affective responses from the Pilot study. Ratings found within the min/max range were deemed as correct (1) and ratings outside the range were deemed as incorrect (0). p(correct) refers to a participant's confidence rating for the corresponding response, noting that the confidence ratings across all trials for a participant were linearly scaled onto a [0,1] range. For example, if a participant were 100% confident with their rating of an advert, this would be transformed to 1. Conversely, if a participant were not confident at all and provided a confidence rating of 0%, this would be transformed to 0. QSR has a range from 0 to number of trials, n. If confidence is randomly set to 0 or 1, QSR would be n/2. We employed this rule for meta-cognitive ability both because of its intrinsic sense and support in the literature (Carpenter et al., 2019, and references therein); additionally, there were insufficient trials for obtaining a stable estimate of meta-d', as used by Carpenter et al. (2019).

3.3.1.3 Procedure

After giving informed consent, participants responded to some simple demographic and meditation questions. They were then provided with some initial instructions regarding the task. Participants were told they would be presented with a series of adverts and would be asked to provide ratings as to how these made them feel. Replicating the procedure of White et al. (2014), trials were organised into two blocks. One block contained two single and two double rating camera adverts, two single and two double rating insurance adverts, two single and two double rating mobile phone adverts, and finally five single and four double rating neutral furniture adverts. In total, this formed 21 trials per block. The second block contained the same adverts but counterbalanced across single/double ratings (note, this also resulted in a total of five double and four single trials in the neutral furniture adverts in the second block). Overall, the number of trials per block was therefore: (2+2)+ (2+2)+ (2+2)+(5+4)= 21 trials. A trial in the experiment consisted of an advert, presented for five seconds to ensure participants had ample opportunity to view the advert, before a request for a rating appeared (or not, depending on the single/double condition). So, trials were grouped into pairs and presentation order of these pairs was randomized within blocks. Moreover, advert pairs were randomised in subsets of three trials, such that these three trials included single and double rating versions for the same adverts. Once all trials were completed, participants answered the FFMQ and then completed the meta-cognition trials. Participants viewed all the adverts sequentially, providing a rating for how positive/negative the adverts made them feel and confidence ratings for their degree of certainty corresponding to each affective response. Once the main experimental task was finished, participants were then provided with a debrief.

3.3.2 Results

3.3.2.1 Data Screening

A preliminary data processing step was conducted. Participants would be excluded on the basis of the following three criteria: (1) missing data/ responses on more than 15% of the trials, (2) overall completion time of more or less than 3.5 standard deviations from the mean (M = 1964 seconds (~33 minutes); SD = 672 seconds (3.5 SDs = 2352), upper limit of M + 3.5 SDs = 4317 seconds (~72 minutes), lower limit of M– 3.5 SDs = 0, as of course we cannot have negative seconds), and (3) incorrect responses on more than 50% of the trials implementing attention checks. All participants who completed the study passed these criteria. Eleven participants partially completed the survey and were excluded from our data analyses.

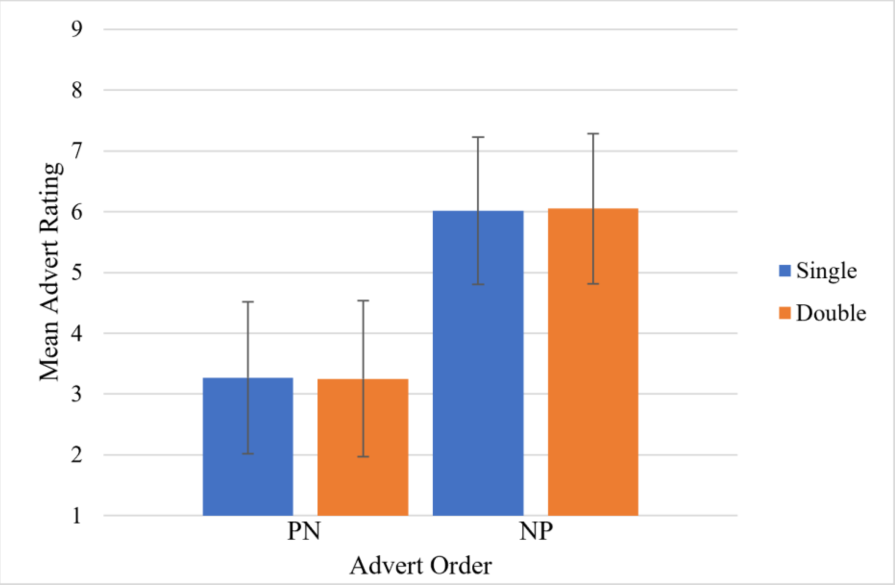
3.3.2.2 Data Analyses

Firstly, we examine evidence for White et al.'s (2014) finding relating to the EB using a within-subjects ANOVA and paired samples t-tests. Then, we analyse the covariance structure between all measures with a series of correlations. To investigate how metacognition and mindfulness mediate the EB, simple linear regressions were built containing the most promising variables based on the pairwise correlations.

3.3.2.3 Replicating White et al. (2014)

We conducted a two (advert order: PN, NP) × two (rating: single, double) repeated measures ANOVA on the ratings for the second adverts. There was a main effect of advert order ($F(1, 200) = 494.64, p < .001$), but not of rating ($F(1, 200) = .107, p = .743$). The key advert order × rating interaction was also not significant ($F(1, 200) = .738, p = .391$); therefore, there was no evidence for an EB in this experiment. Even though this interaction was not significant, for exploratory purposes we offer paired samples t-tests for the main comparisons of interest. Indeed, we found there were also no significant differences in both PN and NP conditions between ratings for the second adverts than those without the intermediate rating (Figure 3.2).

Figure 3.2. Experiment 1 results: mean participant ratings of single and double rated PN and NP adverts (error bars represent ±1 standard deviation).



To sum up so far, we were unable to replicate the effects of White et al. (2014), whereby an intermediate rating can produce stronger affective reactions (whereby the rating of the second advert was more negative in the PN condition and more positive in the NP condition). Instead, we found that affective ratings were similar, regardless of an intermediate rating. Since our second hypothesis was that individual differences in one's ability to self-reflect may impact on the EB, it might be the case that there is an EB for part of the sample which is cancelled out (or undermined) by an opposite effect for the remaining of the sample. We next proceeded to conduct the same ANOVA and paired t-tests but taking into account meditation status.

Meditators versus non-meditators were determined through the recruitment pre-screen in Prolific (*n* = 100 in each) as well as a manipulation check within our demographics, asking whether the participants meditated or not. Contrary to expectation, this check revealed a split of 133 non-meditators and 68 meditators (*n* = 201). There were likely a range of possible explanations for this uneven recruitment of meditators (e.g., participants were meditators when answering their pre-screen, but were not at the time of completing our survey). Unfortunately, the three-way interaction was non-significant, and the corresponding means across the meditators and non-meditators were nearly identical (see Table 3.4 below).

Table 3.4 Descriptive statistics for meditators versus non-meditators in Experiment 1.

	Condition	M	SD
Meditators (*n* = 68)	PN single	3.28	1.49
	PN double	3.38	1.46
	NP single	6.29	1.37
	NP double	6.28	1.27
Non-Meditators (*n* = 133)	PN single	3.26	1.12
	PN double	3.19	1.18
	NP single	5.87	1.1
	NP double	5.93	1.2

Notes: PN = Positive Negative; NP = Negative Positive. Single = rating of second advert (no intermediate rating); Double = rating of second advert (with intermediate rating)

3.3.2.4 Bivariate Correlations Between Measures

Table 3.5 presents the Descriptive statistics for the responses to the facets and overall FFMQ, QSR and EB in this study. Bivariate correlations between the measures of FFMQ, QSR and EB also follow, for the entire sample (see Table 3.6), and split by whether participants were meditators or not (see Tables 3.7 and 3.8).

Table 3.5 Descriptive statistics for the responses to the FFMQ, QSR and EB in Experiment 1.

Measure (Range)	Scale	No. items	M	SD	Skew	Kurtosis
FFMQ (1-5)	Observe	8	27.81	5.39	-.097	-.006
	Describe	8	26.24	6.22	-.306	-.057
	Act with Awareness	8	26.45	6.63	-.106	-.405
	Non-Judgement	8	25.63	7.08	.041	-.507
	Non-React	7	21.52	4.88	-.176	.121
	Overall	39	127.64	19.11	.220	.094
QSR	-	-	15.78	6.86	-.106	-.669

EB	-	-	.21	3.45	.324	2.59
Notes: Standard error for all Skewness is se = 0.17 and for all Kurtosis is se = 0.34.						

The EB was positively related to the NR facet only (*p* = .048), suggesting that returning quickly to an emotional baseline may be related to increased affective responses after rating an intermediate advert. In line with previous research, nearly all correlations between FFMQ facets were significant. The exception here is the relationship between the observe and NJ facets, which was marginally non-significant. Also, the observe facet was significantly negatively related to QSR.

Table 3.6 Bivariate correlations between the measures used in Experiment 1 (*n* = 201).

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.386***	-						
	Act with Awareness	.146*	.424***	-					
	Non-Judgement	-.118^{a}	.213**	.383***	-				
	Non-React	.171*	.234**	.311***	.280***	-			
	Overall	.458***	.720***	.747***	.611***	.591***	-		
QSR	-	-.152*	-.093	-.035	-.006	.003	-.087	-	
EB	-	.089	.023	.084	.047	.139*	.114	-.111	-

Notes: **p* < 0.05, ***p* < .01, ****p* < .001, ^{a} = marginally non-significant: *p* < .1.

We next consider this pattern of correlations separately for meditators and non-meditators, since the two categories of participants may both approach the task differently and may respond to the FFMQ questions in a different way. Within meditators only (see Table 3.7), we observed a stronger relationship between the NR facet and the EB. Similarly, most relationships between the FFMQ facets increased in strength. Within non-meditators only, no relationship was found between the EB and the NR facet. However, there were some marginally non-significant results between the EB and FFMQ, and the EB and QSR (both *p* = .09; see Table 3.8).

Table 3.7 Bivariate correlations between the measures used in Experiment 1 ($n = 68$). Meditators only.

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.518***	-						
	Act with Awareness	.285*	.625***	-					
	Non-Judgement	.024	.264*	.484***	-				
	Non-React	.382**	.445***	.494***	.252*	-			
	Overall	.592***	.802***	.833***	.603***	.706***	-		
QSR	-	-.176	-.025	-.019	-.026	.038	-.056	-	
EB	-	.194	.115	.021	-.143	.241*	.105	-.068	-

Notes: *$p < 0.05$, **$p < .01$, ***$p < .001$.

Table 3.8 Bivariate correlations between the measures used in Experiment 1 ($n = 133$). Non-meditators only.

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.307***	-						
	Act with Awareness	.059	.316***	-					
	Non-Judgement	-.18*	.195*	.337***	-				
	Non-React	-.003	.094	.19*	.311***	-			
	Overall	.349***	.672***	.697***	.641***	.492***	-		
QSR	-	-.124	-.129	-.042	.002	-.014	-.101	-	
EB	-	.086	-.012	.125	.128	.099	.147^{a}	-.148^{a}	-

Notes: *$p < 0.05$, ***$p < .001$, ^{a} = marginally non-significant: $p < .1$.

3.3.2.5 Regression Analyses

Linear regressions were used to predict the EB from the key measures (all individual FFMQ facets, excluding the overall measure due to collinearity, and the QSR variable) in the study. Regressions were conducted first on the entire sample and then separately for whether participants meditated or not. We tested separately for meditators and non-meditators, since the two categories of participants may both approach the task differently and may respond to the FFMQ measures in different ways. Additionally, we also computed a meditation experience variable which goes beyond the pre-screening facility offered at Prolific Academic, which we call the multiplicative meditation variable, MMV. This variable was computed by multiplying the meditation questions by one another (frequency x duration x experience). We computed interaction terms by multiplying the MMV by all the FFMQ variables. These interaction terms were included in each of the following models. After conducting these analyses, we then ran further regressions using only the overall FFMQ measure, QSR and the MMV.

In Model 1, all participants were included, and EB was predicted based on all variables (all FFMQ indices, QSR, the meditation experience variable, and interaction terms between the meditation variable and FFMQ indices). This model was found to be just below the conventional significance level ($F(13, 187) = 1.768$, $p = .051$, $R^2 = .109$). In Model 2, we pursued a reduced model, removing variables producing non-significant correlations with the EB. Therefore, the only variable retained was the NR facet ($F(1, 199) = 3.896$, $r = .139$, $p =.05$, $R^2 = .019$). To reiterate, NR refers to the active detachment from negative thoughts and emotions so that we can accept them and choose not to react to them. Note, we might expect that the non-react facet would be negatively related to the EB since high scorers on the NR facet may return to an emotional baseline quicker, in turn, reducing the EB. However, this was not the case.

We next followed a similar analytical approach, but separately for meditators and non-meditators, determined by meditation status by our Prolific Academic split. First, we turn to the meditator sample. Once again, all variables were entered into a linear regression to predict the EB, including the MMV and the interaction terms. This model explained 33.5% of the variance in the EB ($F(13, 54) = 2.089$, $p = .03$, $R^2 = .335$). After removing all variables correlating non-significantly with the EB, we retained the NR facet (*r* = .24, *p* = .048), the NR interaction term (*r* = .31, *p* = .01) and the MMV only (*r* = .32, *p* = .008). We then created a new model with these variables, which accounted for 14% of the variance in the EB (*F*(3, 64) = 3.484, *p* = .021, R^2 = .14). Critically, experienced meditators who scored higher on the NR facet were more likely to exhibit the EB.

Then, we examined the non-meditator part of the sample. All measures, except for the MMV and interaction terms (since non-meditators did not answer the meditation questions), were entered into a linear regression model. However, this model failed to reach significance (*F*(6, 126) = 1.518, *p* = .18, R^2 = .067). All non-meditator correlations (and standardised betas) with the EB were also non-significant. As such, no ‘reduced’ model (with fewer predictors) was produced for non-meditators.

Finally, we conducted three regression models, one for each of the entire sample, the meditators, and the non-meditators, using only three predictors: FFMQ, QSR and the MMV. The latter was not included in the regression for non-meditators, since these participants did not answer the corresponding questions. All three models were non-significant, respectively (*F*(3, 197) = 1.632, *p* = .183, R^2 = .024; *F*(3, 64) = 2.637, *p* = .057, R^2 = .11; *F*(2, 130) = 2.671, *p* = .073, R^2 = .039.

3.3.3 Discussion

This experiment was unable to replicate White et al. (2014)’s results of an EB. With little to no difference between the single and double conditions across all participants, we explored whether the EB could be revealed by parsing meditators and non-meditators. However, this was also not the case. We then examined the relationships between our main variables of interest, particularly how each facet of the FFMQ and the QSR may relate to the EB in the entire sample, as well as how these relationships may change depending on meditation status. Significant positive correlations were found between the EB and the NR facet (in both the full sample and meditator only split). But, in non-meditators, we found no significant relationships. We then built linear regression models with our variables. The NR facet was found to explain a small amount of the variance of the EB in the full sample (1.9%) and some of the meditator sample (14%; alongside the MMV and NR interaction term). However, all non-meditator correlations were non-significant and so no model was produced for non-meditators.

Why might we have found these results? We offer two reasons for these results. Firstly, it is possible that the experimental protocol was not identical to that of White et al. (2014). To reiterate the procedure, the experiment was programmed so that advert pairs were randomised in subsets of three trials. This meant that these three trials included single and double rating versions for the same adverts. An unintended consequence was that trials with similar adverts would be grouped together and presented within quick succession. This could have reduced any constructive effects, as responding to the same advert in quick succession could have easily introduced response biases, e.g., participants might have explicitly sought to be consistent with their remembered previous judgement. This error was fixed in our replication in Experiment 2. Secondly, we identified an issue with the presentation of the metacognition trials. In this experiment, metacognition trials consisted of viewing an advert, rating how positive/negative the advert is, then finally providing a confidence level of the positive/negative rating. Our reasoning for this was so that the participant could be reminded of the advert whilst providing their confidence rating. Of course, one possible disadvantage of this approach is that participants could check their confidence rating against the image and rating. Conceivably, participants could have even answered the confidence question first. We address this issue in Experiment 2.

3.4 Experiment 2

Experiment 2 was nearly identical to Experiment 1. However, two adjustments were made. We made changes to the randomisation of the stimuli in the White et al. (2014) replication, and the presentation order within the metacognition trials (see materials and procedure for more details).

3.4.1 Method

3.4.1.1 Participants

Two hundred participants were recruited using Prolific Academic (UK nationals only). They were paid £3 for their participation. An equal number of males and females were recruited (*n*

= 100 for each). Participants were between 19 and 69 years old (M_{Age} = 41.34 years old, SD = 12.12; one participant did not disclose their age). As per Experiment 1, participants were asked about their English fluency, with the majority of participants reporting being extremely comfortable in understanding English (*n* = 196), and others reporting feeling less comfortable (*n* = 4). Once again, we recruited equally for meditators and non-meditators using Prolific's pre-screening system. However, only 59 participants reported that they meditate within our survey ($M_{\text{MeditationFrequency}}$ = 5.34 times per week, SD = 5.78; $M_{\text{MeditationDuration}}$ = 18.86 minutes per session, SD = 11.17; the majority (~1SD) of meditators reported having between 9 months and 10 years' experience (see Table 3.9).

Table 3.9 Frequency of meditation experience in Experiment 2 (*n* = 59; *M* = 6.46, SD = 2.97).

Item Number	Meditation Experience	Frequency	Valid Percent	Cumulative Percent
1	1-2 months	4	6.8	6.8
2	3-4 months	6	10.2	16.9
3	5-6 months	4	6.8	23.7
4	7-8 months	2	3.4	27.1
5	9-10 months	2	3.4	30.5
6	11-12 months	4	6.8	37.3
7	1-2 years	13	22	59.3
8	3-4 years	8	13.6	72.9
9	5-10 years	9	15.3	88.1
10	11-20 years	2	3.4	91.5
11	21+ years	5	8.5	100

Notes: Mean and standard deviation reflect item number.

3.4.1.2 Materials and Procedure

Experiment 2 was identical to that of Experiment 1, with the exception of two amendments. First, advert order was fully randomised. Specifically, block presentation was randomised and within blocks trials were randomised across participants. The second amendment relates to the presentation of the metacognition trials. In this experiment, participants viewed the advert and rated how positive/negative it is before clicking onto a new page and providing their confidence rating. This addresses the issue we had in Experiment 1, where participant's metacognitive ratings might have been influenced by the image appearing on the same screen.

3.4.2 Results

3.4.2.1 Data Screening

A preliminary data processing step was conducted. Participants would be excluded on the basis of the following three criteria: (1) missing data/ responses on more than 15% of the trials, (2) overall completion time of more or less than 3.5 standard deviations from the mean (M = 1724 seconds (~29 minutes); SD = 648 seconds (3.5 SDs = 2269), upper limit of M + 3.5 SDs = 3994 seconds, lower limit of M– 3.5 SDs = 0), and (3) incorrect responses on more than 50% of the trials implementing attention checks. Two participants were excluded from our analyses because they exceeded +3.5 SDs in completion time. Three participants partially completed the survey and were also excluded from our analyses.

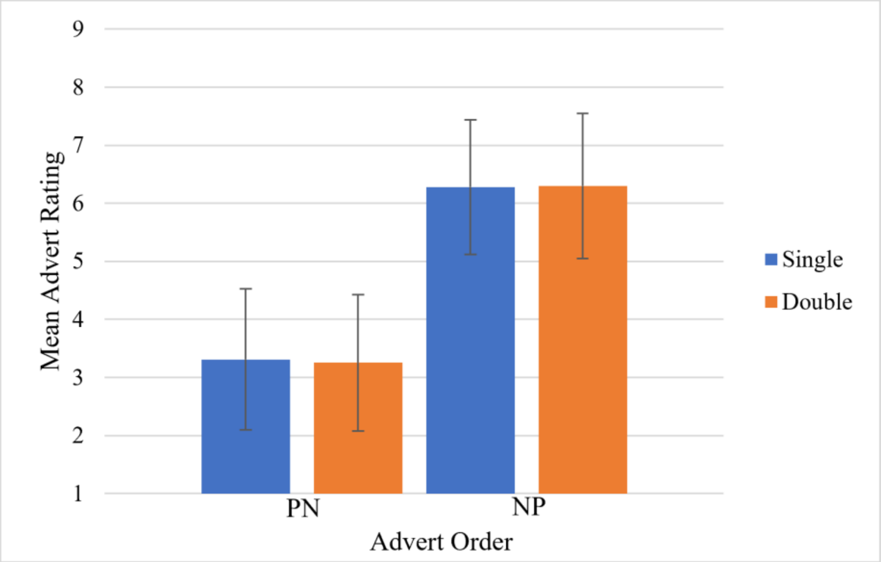
3.4.2.2 Data Analyses

Once again, we are first interested in replicating White et al.'s (2014) finding relating to the EB using a within-subjects ANOVA and paired samples t-tests. We will then analyse all measures with correlations. As per Experiment 1, simple linear regressions will then be built containing variables associated with statistically significant correlations.

3.4.2.3 Replicating White et al. (2014)

We conducted a two (advert order: PN, NP) × two (rating: single, double) repeated measures ANOVA on the ratings for the second adverts. There was a main effect of advert order ($F(1, 199) =558.29, p < .001$), but not of rating ($F(1, 199) = .76, p = .386$). The advert order × rating interaction was marginally non-significant ($F(1, 199) = 3.06, p = .082$). Once again, despite the lack of a significant interaction, we offer paired samples t-tests for the main comparisons of interest. Consistent with the results of Experiment 1, there were no significant differences in both PN and NP conditions for the rating of the second advert, depending on whether the first rated was rated or not (Figure 3.3).

Figure 3.3. Experiment 2 results: mean participant ratings of single and double rated PN and NP adverts (error bars represent ±1 standard deviation).



Once again, we failed to replicate White et al. (2014) regarding the presence of an EB. Regardless of an intermediate rating, affective ratings of the adverts were nearly identical. We next conducted the same analyses as Experiment 1, splitting our sample by meditation status, to test whether mindfulness might interfere with the presence of the EB. Meditators versus non-meditators were determined through the recruitment pre-screen in Prolific (n = 100 in each) as well as a manipulation check within the demographics part of the survey, asking whether the participants meditated or not, how frequent the meditation sessions are (per week), the average duration of a meditation session (in minutes) and how long the participant has been meditating for (months/years). The MMV was once again calculated by multiplying these meditation responses together. Participants who indicated they did not meditate in the survey did not answer the meditation experience questions and were assigned a score of 0 for this variable. This check (that is, splitting participants on the basis of meditation status as in the Prolific Academic recruitment screening) revealed a split of 141 non-meditators and 59 meditators. Meditators scored between 10 and 480 on the MMV variable (*M* = 121.66, SD = 90.03). Given the large range in this meditation score, we reallocated ~10% of the lowest scoring meditators to non-meditators (scores < 30, below 1 SD from the mean), in this way hoping to ensure a cleaner distinction between meditators and non-meditators when we later analyse how meditation status could impact the EB. This procedure resulted in a split of 148 non-meditators and 52 meditators. Note, this approach was not adopted for Experiment 1 since only two participants scored less than 1 SD from the mean in the MMV variable; the range of the MMV variable was narrower in Experiment 1.

For meditators, the three-way interaction was non-significant, and the corresponding means across the meditators and non-meditators were nearly identical (see Table 3.10 below). For non-meditators, there was a significant main effect of advert order (F(1, 147) =417.72, $p <$ .001), but not of rating(F(1, 147) = .292, p = .59). There was also a significant advert order x rating interaction (F(1, 147) = 4.5, p = .036). However, paired samples t-tests revealed no significant differences for the rating of the final advert with and without an intermediate rating (for both NP and PN conditions).

Table 3.10 Descriptive statistics for meditators versus non-meditators in Experiment 2.

	Condition	M	SD
Meditators (n = 51)	PN single	3.32	1.16
	PN double	3.28	1
	NP single	6.06	1.08
	NP double	6.02	1.29
Non-Meditators (n = 149)	PN single	3.31	1.24
	PN double	3.24	1.23
	NP single	6.35	1.18
	NP double	6.39	1.22

Notes: PN = Positive Negative; NP = Negative Positive. Single = rating of second advert (no intermediate rating); Double = rating of second advert (with intermediate rating).

3.4.2.4 Bivariate Correlations Between Measures

Table 3.11 presents the Descriptive statistics for the variables in this study. Bivariate correlations between these measures also follow (see Table 3.12), for the entire sample, and then separately by whether participants were meditators or not (see Tables 3.13 and 3.14).

Table 3.11 Descriptive statistics for the responses to the FFMQ, QSR and EB in Experiment 2 (*n* = 200).

Measure (Range)	Scale	No. items	M	SD	Skew	Kurtosis
FFMQ (1-5)	Observe	8	28.12	5.85	-.296	-.209
	Describe	8	27.88	6.66	-.386	.192
	Act with Awareness	8	27.79	5.99	-.009	-.397
	Non-Judgement	8	27.81	6.94	-.305	-.482
	Non-React	7	20.96	5.06	-.06	.022
	Overall	39	132.54	20.5	.077	.017
QSR*	-	-	15.78	6.61	-.084	-.464
EB**	-	-	.34	2.75	.454	2.848

Notes: Standard error for all Skewness is se = 0.17 and for all Kurtosis is se = 0.34.

We were unable to replicate the results from Experiment 1. Notably, the NR facet was no longer significant in Experiment 2 (*r* = .092, *p* = .197), recall in Experiment 1 we observed *r* = .139, *p* = .048. The only facet to positively relate to the EB was the observe facet. Additionally, the describe and overall FFMQ variables were marginally non-significant (*p* < .1 for both; *r* = .121 and *r* = .138, respectively). As expected, nearly all correlations between FFMQ facets were significant. Also, the QSR was significantly negatively correlated with nearly all FFMQ facets, with the exception of the marginally non-significant NR facet.

Table 3.12 Bivariate correlations between the measures used in Experiment 2 (*n* = 200).

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.465***	-						
	Act with Awareness	.220**	.412***	-					

	Non-Judgement	-.023	.330***	.397***	-				
	Non-React	.296***	.414***	.389***	.245***	-			
	Overall	.566***	.792***	.719***	.615***	.662***	-		
QSR	-	-.259***	-.222**	-.242**	-.162*	-.119[a]	-.301***	-	
EB	-	.154*	.121[a]	-.025	.116	.092	.138[a]	-.062	-

Notes: **p* < 0.05, ***p* < .01, ****p* < .001, [a] = marginally non-significant: *p* < .1.

We once again consider this pattern of correlations separately for meditators and non-meditators. Unsurprisingly, for meditators, the NR facet was highly non-significant in Experiment 2 (*r* = .003, *p* = .98) relative to Experiment 1 (*r* = .241, *p* = .048) for the entire sample. However, meditators showed a significant positive relationship between the describe facet and the EB (see Table 3.13). Relationships between the FFMQ facets were also relatively stable between meditators and the overall sample.

Table 3.13 Bivariate correlations between the measures used in Experiment 2 (*n* = 52). Meditators only.

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.089	-						
	Act with Awareness	.081	.312*	-					
	Non-Judgement	-.107	.497***	.355*	-				
	Non-React	.096	.301*	.436**	.338*	-			
	Overall	.309*	.749***	.672***	.738***	.636***	-		
QSR	-	-.243[a]	-.157	-.257[a]	-.311*	-.115	.011	-	
EB	-	.125	.294*	.052	.155	.003	.219	-.13	-

Notes: **p* < 0.05, ***p* < .01, ****p* < .001, [a] = marginally non-significant: *p* < .1.

Within non-meditators only, no relationship was found between the EB and the describe facet. However, the observe facet was positively related to the EB (see Table 3.14). All other relationships between the EB and the key measures were non-significant.

Table 3.14 Bivariate correlations between the measures used in Experiment 2 (*n* = 148). Non-meditators only.

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.549***	-						
	Act with Awareness	.252**	.442***	-					
	Non-Judgement	-.006	.271**	.411***	-				
	Non-React	.336***	.443***	.378***	.219**	-			
	Overall	.613***	.8***	.734***	.584***	.668***	-		
QSR	-	-.281**	-.254**	-.241**	-.109	-.123	-.297***	-	
EB	-	.192*	.071	-.048	.104	.126	.128	-.032	-

Notes: **p* < 0.05, ***p* < .01, ****p* < .001.

3.4.2.5 Regression Analyses

Linear regressions were then employed to examine how the variables in the study could predict the EB. As in Experiment 1, regressions were conducted first on the entire sample and then separately for whether participants meditated or not. The MMV was calculated and utilised in all regression models as per Experiment 1, except for the models carried out for non-meditators (since these participants would all score 0 on this variable). The MMV was then used to calculate the interaction terms by multiplying it by all other variables in the study.

We first report the models for the entire sample. We attempted to predict the EB based on all variables (the four FFMQ indices, QSR, the MMV, and the interaction variables). The entire sample was used to predict the EB based on the key measures. This model was not found to be statistically significant ($F(13, 186) = 1.234$, $p = .258$, $R^2 = .079$). After removing all variables corresponding to non-significant correlations with EB, we constructed a model with a single predictor, the observe facet ($F(1, 198) = 4.782$, $r = .154$, $p = .03$, $R^2 = .024$). We now explore analogous models for meditators and non-meditators separately.

First, we conducted a linear regression on all key measures to predict the EB for the meditator sample only. However, this model was not statistically significant ($F(13, 38) = 0.839$, $p = .618$, $R^2 = .223$). We then conducted another linear regression retaining only variables correlating significantly with: the describe facet ($r = .29$, $p = .034$), describe interaction term ($r = .32$, $p = .022$) and the MMV ($r = .28$, $p = .045$; $F(3, 48) = 2.801$, $p = .05$, $R^2 = .149$). The describe facet refers to how individuals label their experiences and express them in words to themselves and others. Critically, experienced meditators who score highly on the describe facet also score highly on the EB. Perhaps higher aptitude on describe translates to better articulation of impressions, which leads to a cleaner contrast with the second advert in the double rating condition.

Lastly, we turn to the non-meditator sample. Once again, all measures were entered into a linear regression, and the model was found to be non-significant ($F(13, 134) = 1.406$, $p = .164$, $R^2 = .12$). After removing all variables corresponding to non-significant correlations with EB, we retained only the observe facet ($F(1, 146) = 5.57$, $r = .192$, $p = .02$, $R^2 = .037$), which, recall refers to monitoring and attending to perceptual events. Non-meditators who score higher on the observe facet appear to be more likely to attend to the stimuli shown, responding more strongly to the second stimuli when they have an intermediate rating.

Finally, we conducted three regression analyses, for the entire sample, just meditators, and just non-meditators, using as independent variables the overall FFMQ, QSR and the MMV. All three models were not significant, $F(3, 196) = 1.316$, $p = .27$, $R^2 = .02$; $F(3, 48) = 2.165$, $p = .104$, $R^2 = .119$; $F(3, 144) = 1.077$, $p = .361$, $R^2 = .022$, respectively.

3.4.3 Discussion

Once again, we failed to replicate the EB. We were also unable to replicate correlational results from Experiment 1. In Experiment 1, the NR facet was found to be significant in the full sample and meditator sample. Instead, the observe facet was found to be the only significant facet for the entire sample and non-meditator sample, whereas the describe facet was the only significant facet for the meditator sample. Linear regressions were then created using the respective significant variables. The observe facet was found to explain a tiny amount of the variance in the EB of the full sample (2.4%) and the non-meditator sample (3.7%). However, the describe facet was shown to explain a larger amount of variance in the meditator sample (14.9%).

Given the inconsistent results across Experiments 1, 2, it makes sense to conduct further analyses on the combined samples from the two experiments. In previous experiments, splitting the sample by meditator status resulted in an uneven sample split, with the group of meditators being smaller than the one of non-meditators. We explore this next before outlining our approach to Experiment 3.

3.5 Combined sample

We now consider the combined samples of Experiments 1 and 2. Combining the samples from both experiments allowed us to 'increase' the meditating sample and provide a more reasonable statistical comparison between meditators and non-meditators (see Table 3.15 below).

Table 3.15 Sample split for meditators versus non-meditators in Experiments 1 and 2.

	Meditators	Non-meditators	Total
Experiment 1	68	133	201
Experiment 2	52	148	200
Total	120	281	401

3.5.1 Results

3.5.1.1 Data Screening

Samples from Experiments 1 and 2 were combined. We then conducted a preliminary data processing check on our combined sample of 401 participants. As per the previous experiments, participants would be excluded on the basis of the following three criteria: (1) missing data/ responses on more than 15% of the trials, (2) overall completion time of more or less than 3.5 standard deviations from the mean (*M* = 1827 seconds (~30 minutes); SD = 622 seconds (3.5 SDs = 2177), upper limit of *M* + 3.5 SDs = 4004 seconds (~67 minutes), lower limit of *M* – 3.5 SDs = 0), and (3) incorrect responses on more than 50% of the trials implementing attention checks. We excluded five participants because they exceeded +3.5 SDs in completion time. This resulted in a final sample of 396 participants.

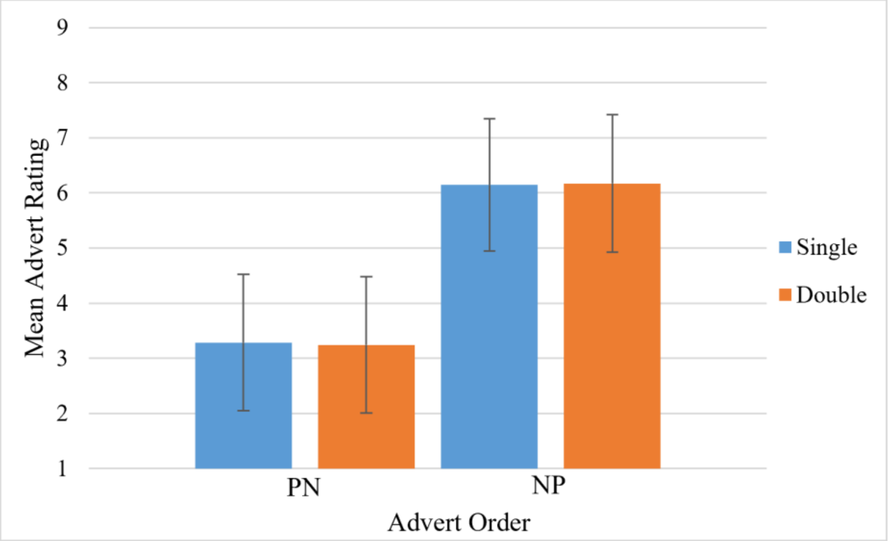
3.5.1.2 Data Analyses

We repeated the same analyses conducted in Experiments 1 and 2. First, we examined whether we could replicate the EB effect from White et al. (2014). Correlations and linear regressions were then conducted on the entire sample, as well as the meditator and non-meditator splits, refining regressions by utilising only the statistically significant correlations of interest.

3.5.1.4 Replicating White et al. (2014)

We conducted a two (advert order: PN, NP) × two (rating: single, double) repeated measures ANOVA on the ratings for the second adverts. Once again, we found a main effect of advert order ($F(1, 395) = 1030.25$, $p < .001$), but not of rating ($F(1, 395) = .16$, $p = .689$). The advert order × rating interaction was marginally non-significant ($F(1, 395) = 3.32$, $p = .069$). Next, we offer paired samples t-tests for the main comparisons of interest, for exploratory purposes given the non-significance of the interaction. Consistent with the results from both experiments, there were no significant differences in both PN and NP conditions for the rating of the second advert, depending on whether the first rated was rated or not (Figure 3.4).

Figure 3.4. Combined Experiments 1 and 2: mean participant ratings of single and double rated PN and NP adverts (error bars represent ±1 standard deviation).



The combined Experiments 1 and 2 failed to replicate White et al.'s (2014) EB effect, showing that, regardless of an intermediate rating, affective ratings of the second adverts were nearly identical. We next split our sample by meditation status, to test whether mindfulness might interfere with the presence of the EB. In both experiments, meditators and non-meditators were determined through the recruitment pre-screens in Prolific (n = 200 in each, 100 for each experiment). We also conducted a manipulation check within the demographics part of the survey, asking whether the participants meditated or not, how frequently participants meditated (per week), the typical duration of their meditation sessions (in minutes) and how long the participant has been meditating for (months/years). We calculated the MMV by multiplying the meditation responses together. Recall, participants who indicated they did not meditate in the survey did not answer the meditation experience questions and were assigned a score of 0 for this variable. This check revealed a split of 271 non-meditators and 125 meditators. Meditators scored between 5 and 770 (*M* = 137.74, SD = 116.31). We reallocated ~7% of the lowest scoring meditators to non-meditators (scores ≤ 20, below 1 SD from the mean). This procedure resulted in a split of 280 non-meditators and 116 meditators.

For meditators, the three-way interaction was non-significant, and the corresponding means across the meditators and non-meditators were nearly identical (see Table 3.16 below). For non-meditators, there was a significant main effect of advert order ($F(1, 279) = 743.62$, $p <$

.001), but not of rating (*F*(1, 279) = .418, *p* = .519). There was also a significant advert order x rating interaction (*F*(1, 279) = 6.69, *p* = .01). Paired samples t-tests revealed an EB in the PN condition: with an intermediate rating, ratings for the second adverts (*M* = 3.21, SD = 1.21) were significantly lower (i.e., the ratings were more negative), than those without the intermediate rating (*M* = 3.29, SD = 1.18; *t*(279) = 2.171, *p* = .031, *d* = .13). However, in the NP condition, no significant differences were found between the second ratings with or without an intermediate rating (*t*(279) = -1.58, *p* = .116).

Table 3.16 Descriptive statistics for meditators versus non-meditators in the combined samples for Experiments 1 and 2.

	Condition	M	SD
Meditators	PN single	3.28	1.37
	PN double	3.31	1.29
	NP single	6.19	1.27
	NP double	6.18	1.29
Non-Meditators	PN single	3.29	1.18
	PN double	3.21	1.21
	NP single	6.12	1.17
	NP double	6.17	1.24

Notes: PN = Positive Negative; NP = Negative Positive. Single = rating of second advert (no intermediate rating); Double = rating of second advert (with intermediate rating).

3.5.1.4 Bivariate Correlations Between Measures

Table 3.17 presents the descriptive statistics for the responses for the variables in this study. Bivariate correlations between these measures also follow (see Table 3.18), for the entire sample and then separately by whether participants were meditators or not (see Tables 3.19 and 3.20).

Table 3.17 Descriptive statistics for the responses to the FFMQ, QSR and EB for the combination of Experiments 1 and 2 (*n* = 396).

Measure (Range)	Scale	No. items	M	SD	Skew	Kurtosis
FFMQ (1-5)	Observe	8	27.96	5.64	-.2	-.15
	Describe	8	27.09	6.51	-.33	.04
	Act with Awareness	8	27.16	6.34	-.1	-.34
	Non-Judgement	8	26.76	7.1	-.13	-.59
	Non-React	7	21.22	5	-.11	.01
	Overall	39	130.18	19.98	.16	.03
QSR*	-	-	15.81	6.71	-.01	-.56
EB**	-	-	.29	3.12	.36	2.93

Notes: Standard error for all Skewness is se = 0.12 and for all Kurtosis is se = 0.25.

Table 3.18 shows the bivariate correlations between the measures from our combined sample. We found that the NR and observe FFMQ facets were significantly positively related to the EB (*p* = .019; *p* = .022, respectively). Note, these facets were individually significant across experiments (NR in Experiment 1 and observe in Experiment 2). The overall FFMQ variable was also significant (*p* = .017). Additionally, the QSR was negatively correlated with the EB, with marginal non-significance (*p* = .067). QSR was also significantly negatively correlated with many of the FFMQ facets (all *p* < .01), excluding NJ and NR. As found in both previous experiments, nearly all correlations between FFMQ facets were highly significant.

Table 3.18 Bivariate correlations between the measures from combined Experiments 1 and 2 (*n* = 396).

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.429***	-						
	Act with Awareness	.182***	.422***	-					

	Non-Judgement	-.065	.282***	.393***	-				
	Non-React	.235***	.32***	.343***	.253***	-			
	Overall	.516***	.761***	.732***	.617***	.619***	-		
QSR	-	-.206***	-.155*	-.133**	-.079	-.055	-.193***	-	
EB	-	.116*	.066	.031	.074	.118*	.12*	-.092[a]	-

Notes: *p < 0.05, **p < .01, ***p < .001, [a] = marginally non-significant: p < .1.

Now we consider correlations separately for meditators and non-meditators (see Tables 3.19 and 3.20). For meditators, the only significant correlation found between the EB and the FFMQ measures was that of the describe facet (p = .015). There were also marginally non-significant relationships between the EB and the NR facet and overall FFMQ variable (p = .096, p = .082, respectively).

Table 3.19 Bivariate correlations between the measures used in the combination of Experiments 1 and 2 (n = 116). Meditators only.

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.354***	-						
	Act with Awareness	.221*	.518***	-					
	Non-Judgement	-.015	.387***	.44***	-				
	Non-React	.273**	.345***	.493***	.253**	-			
	Overall	.495***	.78***	.793***	.661***	.655***	-		
QSR	-	-.225*	-.071	-.111	-.132	-.029	-.167[a]	-	
EB	-	.142	.226*	.042	.014	.155[a]	.162[a]	-.117	-

Notes: *p < 0.05, **p < .01, ***p < .001, [a] = marginally non-significant: p < .1.

In non-meditators, there were two significant findings related to the EB (see Table 3.20). The observe facet was positively related (p = .018) and the overall FFMQ predictor was negatively related to the EB. Interestingly, the NR facet was also marginally significant (p = .053). However, all other relationships between the EB and the other measures were non-significant.

Table 3.20 Bivariate correlations between the measures used in the combination of Experiments 1 and 2 (n = 280). Non-meditators only.

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.448***	-						
	Act with Awareness	.169**	.384***	-					
	Non-Judgement	-.078	.244***	.375***	-				
	Non-React	.201**	.302***	.285***	.258***	-			
	Overall	.513***	.75***	.71***	.608***	.6***	-		
QSR	-	-.207***	-.194**	-.143*	-.055	-.067	-.207***	-	
EB	-	.141*	.011	.03	.096	.116[a]	-.118*	-.081	-

Notes: *$p < 0.05$, **$p < .01$, ***$p < .001$, [a] = marginally non-significant: $p < .1$.

3.5.1.5 Regression Analyses

Linear regressions were then employed to examine how the variables in the study could predict the EB. Once again, regressions were conducted first on the entire sample and then separately for whether participants meditated or not. The MMV was calculated and utilised in all regression models as per Experiments 1 and 2, except for the models carried out for non-meditators (since these participants would all score 0 on this variable). We calculated interaction terms for MMV and the predictors, by multiplying corresponding pairs of variables.

We first offer a model containing all variables for our entire sample, which was not statistically significant (F(13, 382) = 1.841, p = .036, R^2 = .059). We then entered only the variables which had significant correlations with EB into a second model. This model included the observe facet (*r* = .12, *p* = .011), NR facet (*r* = .12, *p* = .009) and QSR (*r* = -.09, *p* = .033). This model was significant and accounted for nearly 3% of the variance in the EB (*F*(13, 392) = 3.607, *p* = .014, R^2 = .027). Next, we consider models for meditators and non-meditators separately.

We first explored the meditator sample and entered all variables into a model. This model was found to be marginally non-significant (*F*(13, 102) = 1.807, *p* = .052, R^2 = .187). Then, we conducted another linear regression retaining only variables correlating significantly with the EB. Thus, the describe facet (*r* = .226, *p* = .007), describe interaction term (*r* = .318, *p* < .001), the NR facet (*r* = .155, *p* = .048), NR interaction term (*r* = .281, *p* = .001), and the MMV (*r* = .279, *p* = .001) were entered into a second meditator model. The second model was significant and accounted for nearly 14% of variance in the EB (*F*(5, 110) = 3.481, *p* = .006, R^2 = .137).

We now turn to the non-meditator sample. Once again, all measures were entered into a linear regression, and the model was found to be non-significant (*F*(13, 266) = 1.524, *p* = .108, R^2 = .069). After removing all variables corresponding to non-significant correlations with EB, we retained only the observe facet (*r* = .141, *p* = .009) and the NR facet (*r* = .116, *p* = .0260. This refined non-meditating model was also significant (*F*(2, 277) = 3.964, *p* = .02, R^2 = .028).

Lastly, we conducted regression analyses separately for the entire sample, meditators only and non-meditators only. Replicating our previous analyses, we conducted these regressions using only the overall FFMQ, QSR and the MMV variables. The entire sample model was significant and explained only 2% of variance in the EB (*F*(3, 392) = 2.659, *p* = .048, R^2 = .02). The meditator model was also significant, accounting for over 11% of the EB (*F*(3, 112) = 4.649, *p* = .004, R^2 = .111). The non-meditating model failed to reach conventional significance (*F*(3, 276) = 2.107, *p* = .1, R^2 = .022).

3.5.2 Discussion

Even when combining Experiments 1 and 2, we were unable to replicate White et al.'s (2014 result. We did find one exception where we found evidence for an EB in the PN condition for non-meditators. However, due to the large non-meditator sample and the tiny difference between the positive and negative stimuli, we do not explore this any further.

So why were we unable to replicate the EB in our experiments? This is a complex question that requires careful consideration. First and foremost is that we consider the possibility that the EB does not exist. This seems the obvious place to start since our data suggests little to no constructive influences across single and double conditions. What is striking is that there have been several empirical demonstrations of the EB across multiple contexts, including the trustworthiness of celebrities and organisation strategies of businesses (White et al., 2016; 2020). This, of course, does not answer why we were unable to find the EB, but may suggest that something within our protocol is amiss. As such, we next consider this point directly.

Second, it is possible that we did not fully replicate White et al. (2014) because of deviation in our methodology. Crucial differences include the experience of the participants recruited for the study, the crowdsourcing platform used, and the stimuli rated in each of the studies. It is exactly this issue that we further explore in Experiment 3.

3.6 Experiment 3

Experiment 3 was conducted to more closely replicate White et al.'s (2014) second experiment and so two amendments were implemented. First, we recruited a sample consisting of participants who would be more experimentally naïve (undergraduate students at the Department of Psychology, City, University of London, instead of participants recruited through Prolific Academic). It is possible that extensive experience with behavioural experiments may have suppressed the key results of interest. Additionally, it is possible that there is a correlation between age and efficacy of mindfulness and so restricting sampling to undergraduate participants has some advantages (see shortly). Second, some of the advertisements in Experiment 3 were reverted back to White et al.'s original adverts. Briefly, in this experiment we employed again hypothetical ads for the 'Blackberry' smartphones, instead of 'Apple' ones. There is a possibility that the latter, being extensively familiar to participants, might be associated with more crystalised emotional responses – and so a bias like the EB might be less likely to emerge. Both amendments are discussed further in materials and procedure.

3.6.1 Method

3.6.1.1 Participants

108 undergraduates were recruited from City University London's participant pool and were given course credit for their participation. We retained a final sample of 95 participants after screening for partial completions and excessively high completion times. All but six of the participants were female (reflecting undergraduate patterns in this field) and were between 18 and 46 years old (M_{Age} = 19.18 years old, SD = 3.21). Once again, participants were asked about their English fluency, with the majority of participants reporting being extremely comfortable in understanding English (*n* = 81), some participants reporting being somewhat comfortable (*n* = 9), and others reporting feeling less comfortable (*n* = 5). Although we could not specifically utilise a pre-screen to determine meditators and non-meditators in our undergraduate sample, participants were still asked whether they meditated, and were asked to report their meditation experience, frequency and duration. In fact, 14 participants reported that they actually meditate within our survey ($M_{\text{MeditationFrequency}}$ = 7.5 times per week, SD = 9.95; $M_{\text{MeditationDuration}}$ = 32.79 minutes per session, SD = 50.06; the majority (~1SD) of meditators reported having between 5 months and 10 years' experience (see Table 3.21).

Table 3.21 Frequency of meditation experience in Experiment 3 (*n* = 14; *M* = 6.07, SD = 2.62).

Item Number	Meditation Experience	Frequency	Valid Percent	Cumulative Percent
1	1-2 months	0	0	0
2	3-4 months	2	14.3	14.3
3	5-6 months	1	7.1	21.4
4	7-8 months	2	14.3	35.7
5	9-10 months	0	0	35.7
6	11-12 months	1	7.1	42.9
7	1-2 years	3	21.4	64.3
8	3-4 years	3	21.4	85.7
9	5-10 years	1	7.1	92.9
10	11-20 years	1	7.1	100

11	21+ years	0	0	100
Notes: Mean and standard deviation reflect item number.				

3.6.1.2 Materials and Procedure

The procedure of Experiment 3 was nearly identical to that of Experiments 1 and 2. However, two amendments were made. We noted large differences between the age of our participants from previous experiments (Experiment 1 M_{age} = 39.97 years old, SD = 12.56; Experiment 2 M_{age} = 41.34 years old, SD = 12.12) and that of White et al.'s (M_{age} = 20.1 years old, SD not available). Previous research has suggested that as a person ages they become more mindful (Hohaus & Spark, 2013). Indeed, both previous experiments also showed that as age increases so do mindfulness scores (see Table 3.22).

Table 3.22 Bivariate correlations between age and FFMQ measures in Experiments 1 (*n* = 199) and 2 (*n* = 199).

Measure	Domain	Experiment 1 Age	Experiment 2 Age
FFMQ	Observe	.091	.076
	Describe	.220**	.096
	Act with Awareness	.206**	.093
	Non-Judgement	.202**	.210**
	Non-React	.047	.095
	Overall	.255***	.175*

Notes: *$p < 0.05$, **$p < .01$, ***$p < .001$.

Additionally, Shook et al. (2017, p.338) contend that "older adults… are more likely to experience negative life events than younger adults, but they tend to have better emotional well-being than younger adults." Taken together, the evidence above may suggest that the older age of our participants in Experiments 1 and 2 (relative to White et al.'s) could have reduced the EB, if we further assume that higher mindfulness leads to lower EB. Either way, seeing that Experiments 1, 2 did not replicate earlier results, since White et al.'s original experiment utilised undergraduate participants, we opted to replicate this aspect of their procedure for Experiment 3. Note, a consequence of recruiting from an undergraduate population was that it was more difficult to specifically recruit for meditators (although, as mentioned above, we nevertheless requested this information from our participants). Moreover, as noted, since the EB is a subtle behavioural bias, participants with extensive experience completing psychology experiments may be less prone to display it. Therefore, we advertised the study at a time such that for many students this would be the first psychology experiment they would ever do.

Our second amendment also closely replicated White et al.'s experiment, since we reverted the 'Apple' advertisement back to White et al.'s original 'Blackberry'. Our reasoning for this change is twofold. First, we aimed to replicate White et al.'s second experiment as closely as possible. Second, as noted, using a brand which participants are likely extensively familiar with, such as 'Apple', may have resulted in a reduced EB, because for such a case there may be crystalised emotional reactions.

3.6.2 Results

3.6.2.1 Data Screening

A preliminary data processing step was conducted. Participants would be excluded based on the following three criteria: (1) missing data/ responses on more than 15% of the trials, (2) overall completion time of more or less than 3.5 standard deviations from the mean (M = 1724 seconds (~29 minutes); SD = 648 seconds (3.5 SDs = 2269), upper limit of M + 3.5 SDs = 3994 seconds (~67 minutes), lower limit of M– 3.5 SDs = 0), and (3) age restrictions on the survey. In total, thirteen participants were excluded from our analyses because they exceeded +3.5 SDs in completion time, partially completed the study or reported an age younger than eighteen.

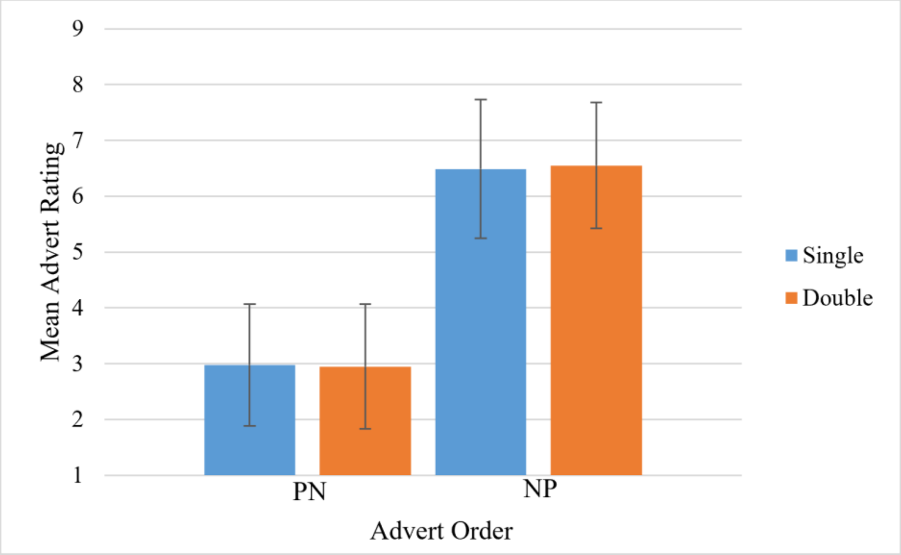
3.6.2.2 Data Analyses

As per Experiments 1 and 2, we will first attempt to replicate White et al.'s (2014) EB using a within-subjects ANOVA and paired samples t-tests. Correlations between all measures will then be examined. Finally, simple linear regressions will be built, via statistically significant correlations, to explore whether metacognition and mindfulness affect the EB.

3.6.2.3 Replicating White et al. (2014)

We conducted a two (advert order: PN, NP) × two (rating: single, double) repeated measures ANOVA on the ratings for the second adverts. There was a main effect of advert order (*F*(1, 94) =364.58, $p < .001$), but not of rating (*F*(1, 94) = .127, $p = .722$). The advert order × rating interaction was non-significant (*F*(1, 94) = 1.01, p = .316). We once again offer paired samples t-tests for the main comparisons of interest, simply for illustration (given the non-significance of the interaction of interest). Consistent with the results of both previous experiments, there were no significant differences in both PN and NP conditions for the rating of the second advert, depending on whether the first rated was rated or not (Figure 3.5.)

Figure 3.5. Experiment 3 results: mean participant ratings of single and double rated PN and NP adverts (error bars represent ±1 standard deviation).



Despite expectation, we failed once more to replicate White et al.'s (2014) EB. Although we may note small rating differences, affective ratings of the adverts were virtually identical across the conditions of interest. We then conducted the same analyses as Experiments 1 and 2, splitting our sample by meditation status, to test whether mindfulness impacts the EB. Meditators versus non-meditators were determined through a manipulation check within the demographics part of the survey, asking whether the participants meditated or not, how frequent the meditation sessions are (per week), the average duration of a meditation session (in minutes) and how long the participant has been meditating for (months/years). In Experiment 3, participants who indicated they did not meditate in the survey also answered the meditation experience questions (in previous iterations, participants did not respond to these). As per our previous experiments, we computed the MMV by multiplying the meditation responses together, as an aggregate measure of meditation experience. We note a possible limitation in this approach in that although non-meditators answered the meditation questions, a 'zero' response on meditation duration renders the MMV score as zero in total, even if responses to the frequency and experience questions are non-zero. The meditator question revealed a split of 81 non-meditators and 14 meditators, whilst the MMV offered a shift in the number of meditators and non-meditators (16 and 79, respectively). MMV meditators scored between 20 and 8000 on the MMV variable (*M* = 1097.19, SD = 1953.8). We opted for the MMV since it offered a more balanced meditator sample (albeit slightly).

For meditators, there was a significant main effect of advert order (*F*(1, 15) =85.631, *p* < .001), but not of rating (*F*(1, 15) = .431, *p* = .522). There was no significant rating x advert interaction (*F*(1, 15) = 1.99, *p* = .178), and this essentially shows a lack of EB. Similarly, for non-meditators, there was a significant main effect of advert order (*F*(1, 78) =282.61, *p* < .001), no main effect of rating (*F*(1, 78) = .001, *p* = .973), and no significant advert order x rating interaction (*F*(1, 78) = .185, *p* = .668).

Paired samples t-tests revealed no significant differences for meditators or non-meditators when considering the rating of the final advert, with and without an intermediate rating (for both NP and PN conditions). These results are unsurprising when considering the descriptive statistics for both meditators and non-meditators (see Table 3.23 below). One might have expected to find significant differences between the meditator's NP single and double conditions, but this was not the case (*t*(1, 15) = -1.28, *p* = .219); note, in this case the sample size was of course limited.

Table 3.23 Descriptive statistics for meditators versus non-meditators in Experiment 3.

	Condition	M	SD
	PN single	2.77	1.3

Meditators (*n* = 16)	PN double	2.69	1.31
	NP single	6.42	1.3
	NP double	6.67	1.22
Non-Meditators (*n* = 79)	PN single	3.01	1.05
	PN double	3	1.07
	NP single	6.5	1.25
	NP double	6.53	1.12

Notes: PN = Positive Negative; NP = Negative Positive. Single = rating of second advert (no intermediate rating); Double = rating of second advert (with intermediate rating).

3.6.2.4 Bivariate Correlations Between Measures

We now turn to the correlations between each of the measures used in Experiment 3. Table 3.24 presents the Descriptive statistics for the responses for the variables in this study. Table 3.25 shows the bivariate correlations between our measures, initially for the entire sample and then separately by whether participants were meditators or not (see Tables 3.26 and 3.27).

Table 3.24 Descriptive statistics for the responses to the FFMQ, QSR and EB in Experiment 3 (*n* = 95).

Measure (Range)	Scale	No. items	M	SD	Skew	Kurtosis
FFMQ (1-5)	Observe	8	26.86	5.52	-.06	-.36
	Describe	8	24.74	6.02	.26	-.14
	Act with Awareness	8	25.33	5.94	-.19	-.42
	Non-Judgement	8	23.99	7.21	-.17	-.63
	Non-React	7	20.24	4.78	.17	.37
	Overall	39	121.16	17.39	.13	.16
QSR	-	-	15.48	6.97	.03	-.67

EB	-	-	.36	3.46	.91	2.46
Notes: Standard error for all Skewness is se = 0.247 and for all Kurtosis is se = 0.49.						

Experiment 3 was unable to replicate the findings of Experiments 1 and 2. Neither the NR facet nor the observe facet were significantly correlated with the EB measure (*r* = .03, *p* = .787; *r* = .05, *p* = .65, respectively); recall in Experiment 1 we observed *r* = .139, *p* = .05 for NR and, in Experiment 2, we observed *r* = .15, *p* = .029 for the observe facet. In fact, all measures in Experiment 3 were non-significant, however QSR was marginally non-significant (*r* = .17, *p* = .095). Once again, we found that nearly all correlations between the FFMQ facets were significant. Additionally, QSR was significantly negatively correlated with the describe facet only.

Table 3.25 Bivariate correlations between the measures used in Experiment 3 (*n* = 95).

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.085	-						
	Act with Awareness	.065	.488***	-					
	Non-Judgement	-.286**	.28**	.389***	-				
	Non-React	.304**	.28**	.174$_a$	.055	-			
	Overall	.333**	.733***	.74***	.569***	.551***	-		
QSR	-	.112	-.294**	-.048	-.139	-.09	-.165	-	
EB	-	.047	-.06	.022	-.012	.028	·.004	.173$_a$	-

Notes: ***p* < .01, ****p* < .001, ^{a} = marginally non-significant: *p* < .1.

We now turn to the correlations for meditators and non-meditators separately. For meditators, there were no significant relationships between the EB and all other measures. This was also the case for QSR. Interestingly, many relationships between the FFMQ facets also seemed to break down (resulting in non-significance or marginal non-significance), but this is likely explained by the smaller sample size.

Table 3.26 Bivariate correlations between the measures used in Experiment 3 (*n* = 16). Meditators only.

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.105	-						
	Act with Awareness	.279	.338	-					
	Non-Judgement	-.06	.217	.575*	-				
	Non-React	.332	$.495_a$	.348	.03	-			
	Overall	.552*	.594*	.828***	.605*	.602*	-		
QSR	-	.039	.026	.327	.198	.241	.262	-	
EB	-	-.099	.211	.098	.188	-.246	.063	.232	-

Notes: *$p < 0.05$, ***$p < .001$, ^{a} = marginally non-significant: $p < .1$.

For non-meditators, we observe a similar pattern. No relationships were found between the EB and all other measures. Once again, we find similar results for QSR, except that QSR was significantly negatively related to the describe facet and overall FFMQ variable, and marginally non-significant with the NJ facet. Relationships between the FFMQ facets were relatively stable, with few unexpected non-significant findings relating to the observe and NR facets.

Table 3.27 Bivariate correlations between the measures used in Experiment 3 (*n* = 79). Non-meditators only.

Measure	Domain	Observe	Describe	Act with Awareness	Non-Judgement	Non-React	Overall FFMQ	QSR	EB
FFMQ	Observe	-							
	Describe	.048	-						
	Act with Awareness	.000	.514***	-					
	Non-Judgement	-.348**	.29*	.357**	-				
	Non-React	.291**	.244*	.142	.056	-			

	Overall	.255*	.749***	.726***	.57***	.538***	-		
QSR	-	.145	-.339**	-.127	-.2$_a$	-.139	-.249*	-	
EB	-	.06	-.13	-.003	-.051	.059	-.035	.169	-
Notes: **p* < 0.05, ***p* < .01, ****p* < .001, ^{a} = marginally non-significant: *p* < .1.									

3.6.2.5 Regression Analyses

We utilised linear regressions to investigate whether the variables in the study could predict the EB. Once again, these regressions were conducted first on the entire sample and then separately depending on whether participants meditated or not. The MMV was calculated and applied in all regression models. Recall, MMV is the multiplicative meditation variable which was computed by multiplying the meditation questions (frequency x duration x experience). This included some non-meditators since, unlike Experiments 1 and 2, these participants could score above 0. The MMV was then used to calculate the interaction terms by multiplying it by all other variables in the study.

We now turn to the models for the full sample. Consistent with our previous experiments, we included all variables (five individual FFMQ indices, QSR, MMV and interaction variables) in a regression to predict the EB. This model was not found to be statistically significant (*F*(13, 81) = .904, *p* = .552, R^2 = .127). After removing all variables corresponding to non-significant correlations with EB, we constructed a model with a single predictor, QSR (*r* = .173, *p* =.047). We now explore analogous models for meditators and non-meditators separately.

First, we conducted a linear regression on all key measures to predict the EB for the meditator sample only. However, this model was not statistically significant (*F*(12, 3) = 0.95, *p* = .595, R^2 = .79). We conducted no further regressions with meditators since there were no significant relationships between any of the variables and the EB.

Second, we turn to the non-meditator sample. All measures were entered into a linear regression, and the model was found to be non-significant (*F*(6, 72) = .625, *p* = .71, R^2 = .05). All variables were found to be non-significant with the EB, with QSR just failing to reach conventional significance (*r* = .169, *p* = .068).

Lastly, we turn to the final regression analyses. The overall FFMQ, QSR and the MMV variables were all entered a regression for the entire sample, just meditators, and just non-meditators separately. All three models were not significant, *F*(3, 91) = 1.19, *p* = .318, R^2 = .038; *F*(3, 12) = .253, *p* = .858, R^2 = .059; *F*(2, 76) = 1.122, *p* = .331, R^2 = .029, respectively.

3.6.3 Discussion

Despite more closely replicating White et al.'s (2014) methodology, we were unable to find the EB. Additionally, there were also no significant relationships between the EB and the key variables in this study. Since we were unable to replicate this result, we conduct no further experiments. Instead, we consider an alternative approach to analyse constructive influences.

3.7 Evaluation Bias Discussion

In a series of three experiments, we examined whether introspection (via metacognition and mindfulness) impacted an individual’s propensity for constructive influences in evaluative decision making. We were interested in how the EB might relate to such individual characteristics. Specifically, we posited multiple lines of reasoning for how a mindful and non-mindful person may elicit the EB. For example, highly mindful individuals may have increased awareness of a stimulus, which may facilitate stronger ratings, or, alternatively, they may be desensitised by such stimuli, resulting in a similar rating and reduce the chances of an EB from occurring. We also offered similar reasoning for how different levels of metacognition could produce the EB. Notably, a person with good metacognition (high confidence when correct and low confidence when incorrect) could be more aware of a stimulus than a person with poor metacognition (low confidence when correct and high confidence when incorrect). A greater awareness of the stimuli may result in stronger ratings as these individuals will be more able to easily recall the previous stimuli to memory, which in turn would enable the EB to occur.

However, our results revealed that we were unable to replicate the EB, as demonstrated by White et al. (2014). It is important to note that we were able to find an EB effect in the combined experiments of 1 and 2, where non-meditators showed an EB in the PN condition. Indeed, the minor difference (a .08 difference on a scale ranging from 1 to 9, see 3.16) between them may simply be explained by the large sample we obtained by merging the experiments. Before exploring why we were unable to replicate the EB, we will briefly revisit why we expected to find it.

Recall, research has shown that context (e.g., question and evidence order) impacts how an individual formulates a response. These findings imply a constructive effect in the way judgements develop, specifically that articulating an impression impacts future decisions. QT allows a more constrained prediction for what this impact is. White et al. (2014) tested this and found that an intermediate judgement makes the rating of the second stimulus more intense compared with the single rating condition (the EB). Moreover, White and colleagues replicated this finding across different contexts. So why could we not replicate the EB here?

Let us address this question by first exploring the differences between White et al.'s experiment and our replication. An obvious starting point is that perhaps our experiments could not replicate the EB because we did not employ the same stimuli, as was used in White et al.'s original experiment. Recall, our pilot study examined a total of 60 adverts, 36 of which were employed in White et al.'s original experiment; note, White et al. used 48 adverts in Experiment 2. The 12 adverts we did not employ comprised of positive and negative filler stimuli and so should not have impacted the ratings of the stimuli in the single and double conditions, and therefore the EB in White et al.'s results. We also employed 24 neutral adverts, that White et al. did not. But, of course, the ratings for these 24 neutral adverts did not affect the computation of the EB. After conducting the pilot and verifying the valence of the images, we removed 18 adverts because they did not map onto their intended valence (i.e., all removed negative adverts were rated as neutral and were rated as 3.92, 4.45, 4.41 and 4.33). Of those removed, 8 adverts were previously employed as key PN or NP stimuli within White et al.'s experiments. Despite our remaining stimuli mapping onto their intended valence, we did not replicate White et al.'s experiment. Could the removal of some adverts have resulted in us not finding the EB? It is possible for this to have had an impact in reducing the EB if some adverts had a greater evaluative impact than others. For example, we expect a higher evaluative influence for stimuli which are slightly more ambiguous. As White et al. (2020, p.26) note,

"Without some ambiguity, no constructive or [QT] effects are expected. For example, if participants see a hammer and are asked if there is a hammer, no changes to the mental state are really expected. Note that there is a converse point here, namely that with too much ambiguity the representational assumptions embodied in the QT model are challenged…"

So, perhaps, by including adverts which were more clearly negatively or positively valenced, we undermined the chance of obtaining an EB. There were also differences in sampling characteristics, namely that White et al. recruited university students whereas we crowdsourced in Experiments 1 and 2 via Prolific. One consequence of this divergence was differences in sampling age (since older individuals typically exhibit higher levels of mindfulness/introspection). However, after conducting Experiment 3 on undergraduate students, we once again did not obtain an EB.

There are also some important limitations to consider with regards to mindfulness and metacognition, and their associated psychometrics. First, not all meditation is mindfulness meditation. The term 'meditation' refers to a range of different practices, each offering varying 'mindful' benefits. Indeed, some meditation techniques might improve certain types of skills (e.g., attention regulation) but not necessarily others (e.g., experiencing thoughts/feelings without judging them or criticizing oneself). It was unclear from our meditation sample exactly how many participants were practicing mindful meditation. In our experiments, we asked whether participants meditated or not (along with some simple experience related questions, such as frequency and duration of meditation). Second, experience with mindfulness meditation may lead a person to interpret and respond to the items on the FFMQ in a different way compared to non-meditators. Specifically, mindful meditators may be more aware of when they are not acting with awareness. This may not be the case for other types of meditators. Third, simply asking metacognitive questions (e.g., confidence ratings) can result in poorer metacognitive thinking by eliciting less accurate retrospective self-appraisals (see Double & Birney, 2019, for review). Finally, we consider physical and psychological self-awareness, particularly whether present moment awareness can be decoupled into awareness of physical sensations/external cues versus awareness of thoughts/emotions. It is possible these are independent and have quite different effects on behaviour. For example, paying attention to the taste of food as one eats could have quite different effects to paying attention to feelings of hunger or satiety, especially if one is eating delicious food when full. Applying this to the current paradigm, perhaps participants were more self-aware of physical sensations than psychological ones. Recall that participants were asked how positive or negative the advert made them feel. For example, if participants engaging with the negative stimuli considered how the stimuli made them feel physically (such as the extent to which they may cry), this could be intrinsically different to a participant considering how negative they feel psychologically. Indeed, we are unable to rule out this notion because we were not specific enough in our meditation questions.

Recall that the EB shows a constructive influence because we are comparing the response to the same stimulus in the same position with or without a response to the previous stimulus. This is one approach we could take to compute the EB. An alternative approach could examine how the pilot ratings reflect the baseline of valence. This can inform us of how the adverts were rated on an individual basis relative to the ratings in the main experiments. First, we will examine how White et al.'s pilot ratings compare to the ratings in their main experiment (which we attempted to replicate; Table 3.28). We then conducted the same analysis for our results (Table 3.29).

As can be seen in Table 3.28 below, White et al. found that viewing an N stimulus in a PN pair weakened the rating of the negative advert relative to if the negative advert was considered by itself (the rating for the same N stimulus in the pilot). In other words, ratings from the pilot study were more intensely negative than both PN single and double ratings. The rating for the N stimulus in the double condition was more negative than in the single condition. This pattern was also observed for the NP single condition, with the exception for NP double which slightly intensified ratings relative to the pilot.

Table 3.28 Descriptive statistics for Pilot versus Experiment 2 ratings in White et al. (2014). Note, we replicated the procedure of Experiment 2 from White et al. Also, Experiments 1 and 3 in White et al. contained variations of the paradigm unrelated to our analysis, and so were not included.

	Condition	M	SD	CI 95%	Absolute differences between the pilot study and Experiment 2
Pilot study (*n* = 12)	Negative	3.09	.83	.47	-
	Positive	6.53	.89	.5	-
	PN single	4.36	.69	.18	1.27

Experiment 2 (*n* = 54)	PN double	3.81	.71	.19	.72
	NP single	5.6	.75	.2	.93
	NP double	6.63	.73	.19	.1
Notes: PN = Positive Negative; NP = Negative Positive. Single = rating of second advert (no intermediate rating); Double = rating of second advert (with intermediate rating). The M values refer to the rating for the second stimulus in all cases.					

Table 3.29 shows a more complex picture than that of Table 3.28. Means of each of the PN, NP conditions were generally clustered around the pilot ratings, with the greatest difference being .27; note the direction of the mean differences is inconsistent. For instance, our Experiment 1 mirrors the trends seen in White et al.'s Experiment 2 in all but magnitude. Experiment 2 also reveals small increases in positivity in the PN condition, but also small increases in the NP condition (otherwise expected to be lower than the pilot rating). Finally, Experiment 3 reveals intensified ratings for both PN and NP conditions relative to the pilot, although this effect was stronger when a negative advert was rated first.

All this is to say that it is possible that there are constructive effects when showing a P or N stimulus, after a previous one, as evidenced by the differences in these ratings, relative to when the same stimuli were rated individually. If real, such constructive effects would be underwritten by a covert response to the first stimulus, regardless of whether one was requested or not. We stress that we have no direct evidence for such a process. Instead, these analyses potentially reveal another locus of interest, namely the extent to which the first stimulus impacts on the second one.

Table 3.29 Descriptive statistics for Pilot and Experiments within the current paper.

	Condition	M	SD	CI 95%	Absolute mean difference between the pilot study and Experiment 2
Pilot study (*n* = 191)	Negative	3.13	.49	.07	-
	Positive	6.22	.38	.05	-
Experiment 1 (*n* = 201)	PN single	3.27	1.26	.17	.14
	PN double	3.26	1.28	.18	.12

	NP single	6.01	1.2	.17	.21
	NP double	6.05	1.23	.17	.17
Experiment 2 (*n* = 200)	PN single	3.32	1.22	.17	.19
	PN double	3.26	1.17	.16	.13
	NP single	6.28	1.14	.16	.06
	NP double	6.31	1.24	.17	.09
Experiment 3 (*n* = 95)	PN single	2.98	1.09	.22	.15
	PN double	2.95	1.11	.22	.18
	NP single	6.49	1.25	.25	.27
	NP double	6.55	1.13	.23	.33

Notes: PN = Positive Negative; NP = Negative Positive. Single = rating of second advert (no intermediate rating); Double = rating of second advert (with intermediate rating). The M values refer to the rating for the second stimulus in all cases.

3.8 The persistence of positive versus negative information

The EB and the corresponding putative role of mindfulness and metacognition are one possible focus for Experiments 1-3. As we have seen, difficulties in consistently replicating the EB challenge a principled study of the role of metacognition and mindfulness. Concerning these three experiments, there is an alternative analytical and theoretical focus. Recall, participants were asked to view and rate the second stimulus independently of the first. If the difference between the ratings for the same stimulus when it is rated first versus when it is rated second is zero, then that would imply that the second stimulus is evaluated independently of the first (as ostensibly intended). However, if this difference is greater or less than zero, then the evaluation of the second stimulus must depend on the first stimulus as well. That is, the impression from the first stimulus ‘sticks’ to affect the second stimulus as well. This differs from our EB analysis, since the EB concerns the impact of the evaluation itself, that is, it involves comparing the rating for the second stimulus with and without an intermediate rating. With this alternative analysis, we are concerned with carry over effects for a stimulus evaluation in the second position, from the first stimulus. Note, such carry over effects do not specifically require ideas from quantum models – these models motivated the initial examination of the EB, but they have less relevance here.

To conduct such an analysis, we first need to consider the rating of a given stimuli when it is viewed and rated first (i.e., the stimulus is rated independently). Let us call this rating I. We then consider the same stimulus when it is viewed second and rated second and viewed second and rated first (i.e., when the prior stimulus is not rated). Respectively, these scores correspond to the double (SSDR) and single (SSSR) scores we utilized in the analysis for the EB above. For simplicity, we will call these D and S.

Each of the I, S and D scores were computed as follows. I scores were calculated by averaging all stimuli presented and rated first in the double rating condition for each valence type. For example, positive I scores were computed by averaging all positive stimuli presented first in the double rating condition, creating $I_{positive}$. The same process was then conducted for the negative stimuli, to compute $I_{negative}$, and neutral stimuli, to compute $I_{neutral}$. D scores were calculated by averaging across all stimuli presented second in the double rating condition and S scores by averaging across all stimuli when presented second in the single rating condition, again for each valence separately.

Finally, we adjusted the S and D scores, by the I scores, separately for each condition (PN and NP). For the PN condition we computed $S' = S – I_{negative}$ and $D'= D - I_{negative}$; the logic of placing the differences in this way is that all ratings (S, D, $I_{negative}$) concern the same stimulus, which in this condition is a negative stimulus, but when this stimulus is evaluated second, it may be less negative than when first, if the prior positive stimulus affects it in an assimilative way. This makes intuitive sense when we consider a simple example. If a participant rated a particular negative stimulus with a 3 in a $I_{negative}$ trial, 4 in the S trial (it is assumed to be more positive, if there is an assimilation effect) and 5 in the D trial, the S' and D' would be as follows: $S' = 4-3 = 1$, and $D' = 5-3 = 2$. Thus, S' and D' would be positive numbers as long as there is assimilation between the first and the second stimulus in a PN pair.

For the NP condition, using a similar logic, we would have $S'= I_{positive} - S$ and $D'= I_{positive} - D$. It is worth noting that S' and D' will also be positive numbers when S and D are lower than $I_{positive}$, which would be the case again if we have assimilation. In both cases, as a trivial baseline, we also set $I'=0$. Therefore, if the first stimulus does not affect at all the evaluation of the second, $S'=0=D'$. Note, this scheme does not commit to assimilation versus contrast, it just offers a convenient coding convention so that positive numbers for S', D' indicate assimilation and negative numbers contrast.

Unfortunately, the above approach cannot be realised exactly as stated, because in our experiments (mostly replicating White et al., 2014), stimuli were either shown in the first or second position, but not in both positions. The above approach is best adopted when the same stimuli would be presented in both positions and this data is not available in the main experiments. Fortunately, the necessary information is available in the pilot results we obtained concerning the valence of the stimuli. So, $I_{positive}$ and $I_{negative}$ values were determined from the pilot results and the S, D ones from the main Experiments 1-3 above. Table 3.30 presents the corresponding results across all three experiments. Indeed, there is a general trend for a higher impact of N stimuli than P ones when shown first, but this trend is not visible in all instances.

Table 3.30 Rating differences for when the same stimulus is presented second with intermediate ratings (D) and no intermediate ratings (S) between White et al. (2014) and the current pilot. Note, D' and S' are computed relative to I' which, by definition, equals zero.

Experimenter	Exp #	Condition	D'	S'
White et al.*	Experiment 1	NP	1.35	1.69
		PN	.94	1.24
	Experiment 2	NP	-.35	.69
		PN	.71	1.26
Current pilot study**	Experiment 1	NP	-.02	.01
		PN	.09	.11
	Experiment 2	NP	-.26	-.24
		PN	.09	.15
	Experiment 3	NP	-.52	-.46
		PN	-.22	-.19

Notes: PN = Positive Negative; NP = Negative Positive. I = a particular stimulus when presented first; S = same stimulus when presented second, in the single rating condition (no intermediate rating); D = same stimulus when presented second, in the double rating condition (with intermediate rating).

*$I_{negative}$ = 3.1; $I_{positive}$ = 6.29. 'I' was derived from White et al.'s (2014) pilot study.

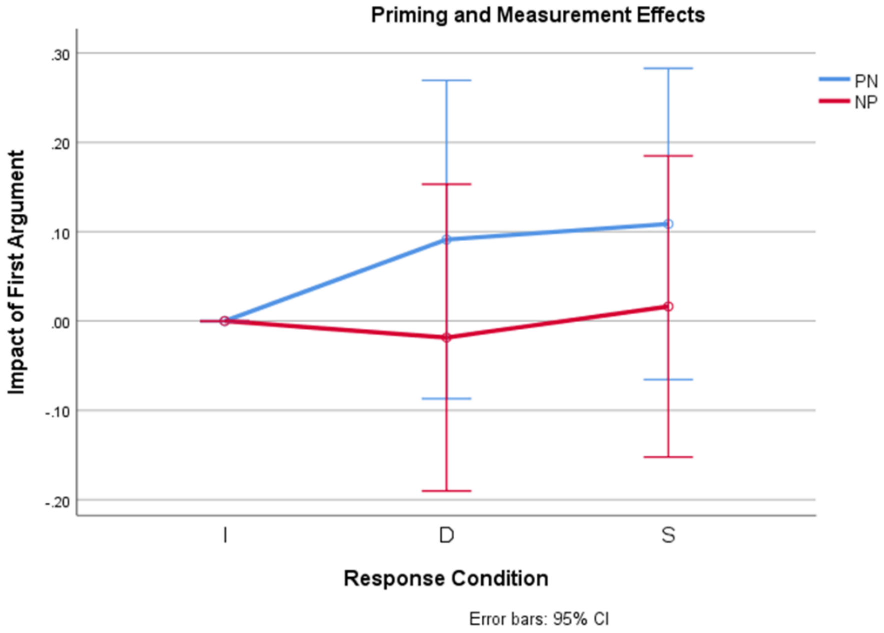
** $I_{negative}$ = 3.16; $I_{positive}$ = 6.03. 'I' was derived from the current pilot study.

We will now explore in more detail the influence of the first stimulus on the second across our three different experiments. Recall that if D and S are different from zero, the evaluation from the first stimulus must influence the second; if the first stimulus does not impact at all on the second, then we expect I = 0 = D' = S'. We will analyse each experiment in turn. Figures 3.6-3.8 illustrate the I, D and S values from Table 3.30.

3.8.1 Experiment 1

We conducted a two (advert order: PN, NP) × two (rating: D, S; note, henceforth we will avoid the primes for simplicity) repeated measures ANOVA on the D'/S' ratings for the stimuli presented in Experiment 1. There was no main effect of advert order or rating or interaction. This suggests that the first stimulus had no evaluative impact on the second stimulus for both advert orders (see Figure 3.6).

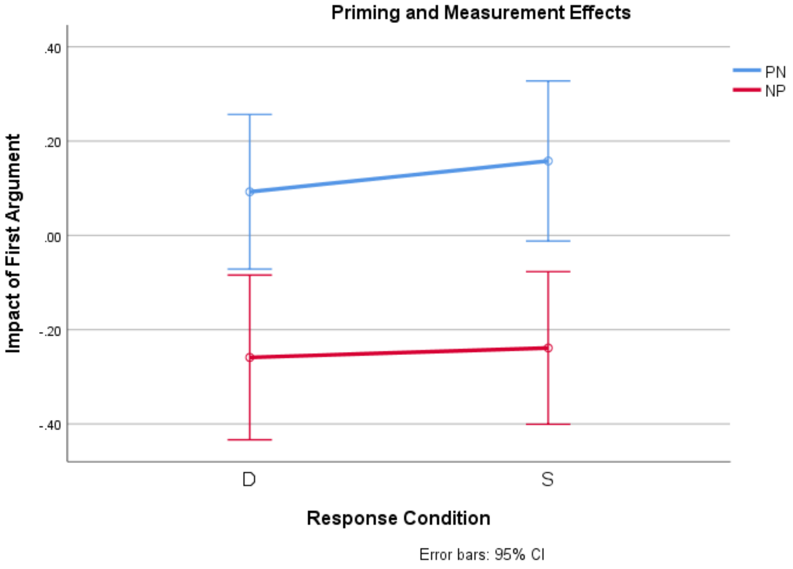
Figure 3.6. Experiment 1: Ratings for stimuli with a certain valence when presented first (I), second with the first stimulus not rated (S), and second with the first stimulus rated (D). The two lines distinguish between the NP and PN orders. Error bars represent 95% confidence intervals.



3.8.2 Experiment 2

We conducted a two (advert order: PN, NP) × two (rating: D, S) repeated measures ANOVA on the ratings for the stimuli presented in Experiment 2. There was a significant main effect for advert order (F(1, 198) = 12.138, p = .001, η^2 = .058), and a non-significant main effect for the rating condition (F(1, 198) = 3.062, p = .082). The advert order × rating interaction was non-significant (F(1, 198) = .849, p = .358; see Figure 3.7.). In this case, it appears that when the first stimulus is negative, it impacts on the rating of the second stimulus more so than if the first stimulus is positive. Recall, in the NP condition e.g. $S' = I_{positive} - S$, so it appears that the same P stimuli when presented second were rated more positively (-.24 for S and -.26 for D) than when presented first; this is a contrast effect (for the NP condition). Interestingly, in the PN condition, the difference indicates that the negative D and S stimuli were being rated more positively in the second position (i.e., following an initial positive stimulus) than when they were in the first position (.15 for S and .09 for D); that is, in the PN condition we appear to have an assimilation effect.

Figure 3.7. Experiment 2: Ratings for stimuli with a certain valence when presented second with the first stimulus not rated (S) and second with the first stimulus rated (D). The two lines distinguish between the NP and PN orders. Error bars represent 95% confidence intervals.



3.8.3 Experiment 3

We conducted a two (advert order: PN, NP) × two (rating: D, S) repeated measures ANOVA on the ratings for the stimuli presented in Experiment 3. There was a main effect for the advert order (F(1, 94) = 5.05, p = .027, η2 = .051), and for the rating condition (F(1, 94) =

1.014, *p* = .316). The advert order × rating interaction was non-significant (*F*(1, 94) = .127, *p* = .722; see Figure 3.8.). Partly replicating Experiment 2, these results indicate that when the first stimulus is negative, it impacts on the rating of the second stimulus more so than if the first stimulus is positive. Indeed, in the NP condition, positive D and S stimuli were rated more positive than their ‘I’ counterparts (-.46 for S and -.52 for D). Conversely, in the PN condition, negative D and S stimuli were rated more negatively than their ‘I’ counterparts (-.19 for S and -.22 for D). So, note, in this case, for both the PN and NP order we appear to obtain contrast effects, instead of assimilation effects.

Figure 3.8. Experiment 3: Ratings for stimuli with a certain valence when presented second with the first stimulus not rated (S) and second with the first stimulus rated (D). The two lines distinguish between the NP and PN orders. Error bars represent 95% confidence intervals.

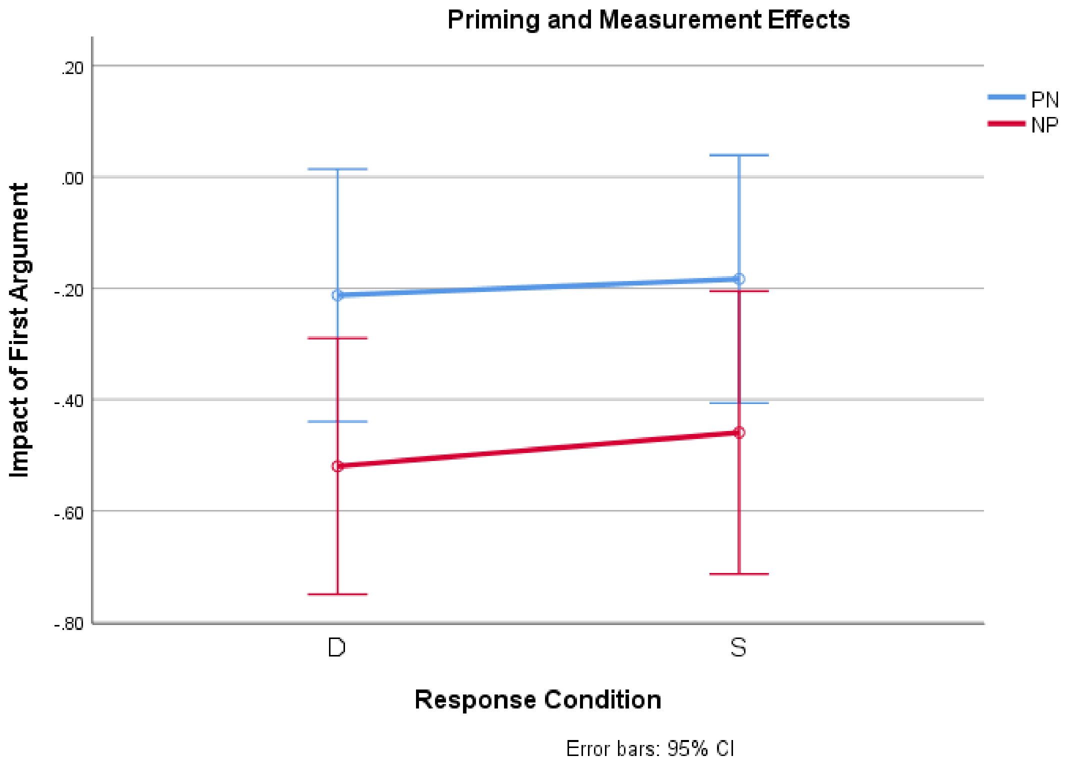


Table 3.31 Rating differences for when a stimulus is presented first versus second in the Pilot study.					
	Condition	First rating	Second rating	Average first position – average second position	Intensity: distance from midpoint*
Pilot	N	3.38	3.16	.22	1.34

	P	6.02	6.03	-.01	1.53

Notes: N = Negative; P = Positive. Each advert in the first position was averaged across the relevant valence (e.g., all positive or all negative), and likewise for those adverts in the second position. However, it is important to note that these stimuli were different adverts across the first and second positions.
*Intensities were computed from the (pilot) ratings of the stimuli appearing in the second position. Note also, ratings were computed by: P – midpoint and midpoint – N.

3.9 General Discussion

Why do negative impressions appear to influence the rating of a subsequent, positive stimulus, more so than the other way round? One model which may explain such effects is called the Inclusion/Exclusion Model (IEM; Schwarz & Bless, 1992; Bless & Schwarz, 2010). This model outlines how context effects arise in feature-based evaluative judgements. In their paper, Schwarz and Bless (1992, p.72) note that:

> "…individuals who are asked to form a judgement about some target stimulus first need to retrieve some cognitive representation of it. In addition, they need to determine some standard of comparison to evaluate the stimulus. Both, the representation of the target stimulus and the representation of the standard are, in part, context dependent. Individuals do not retrieve all knowledge that may bear on the stimulus, nor do they retrieve and use all knowledge that may potentially be relevant to constructing its alternative. Rather, they rely on the subset of potentially relevant information that is most accessible at the time of judgement."

According to IEM, depending on how impressions are processed, assimilation and contrast effects can emerge. The assimilation effect can occur when the initial information is used to form a temporary representation of a target. If a subsequent judgement is made based on this target, an introduction of positive information could result in a more positive judgement. Schwarz and Bless (1992) found this when they asked participants about a popular German politician, prior to asking about the political party he was affiliated with. In other words, simply asking a question about the well-respected politician resulted in a more favourable evaluation of the party compared to if no question was asked at all. Note, the assimilation effect also occurs when negative information is introduced, rendering a more negative judgement.

However, when information is used to form a representation of a standard to which the target is compared, contrast effects are assumed to occur. Schwarz and Bless (1992) demonstrated this effect with the same politician used in the assimilation example above. Since this politician also served as President of the Federal Republic of Germany (a position known to require neutrality and to be separated from party politics), the authors utilised this situation to ask other participants if they knew what position he held which set him aside from party politics. Schwarz and Bless (1992) compared the position of President to the Queen of England, notably that they are both heads of state who do not 'take sides' in parliament. This question was therefore designed to 'remove' any impressions about the politician from assumption that he/she belongs to a particular party, from the minds of participants who were later asked for an evaluation of the political party. That is, since the politician is neutral as head of state, he/she would no longer stand for the party and thus would be removed from a participant's representation of the party. Schwarz and Bless (1992) showed that participants who were asked about the politician's presidency evaluated the party less favourably than the control (which did not include a question about the politician).

The authors offered two explanations for these effects – assimilation when the mention of the politician is 'consistent' with the evaluation of the party and contrast when the mention of the politician is meant to be 'discounted' from the evaluation of the party. First, control participants may have recalled the politician into their mind when evaluating the party (unintentionally eliciting assimilation effects), which may have resulted in a more positive impression of the party, in the assimilation condition. In the contrast condition, participants who considered the presidency of the politician removed him from their cognitive representation of the political party. Accordingly, these participants utilised the politician as a baseline of favourable impressions relative to the party in question, creating a contrast between the standard (the popular politician) and the target (the political party minus the popular politician). Pertinent to this discussion is the extent to which a participant is aware of the first stimulus they encountered, and how it later influences their evaluation of the second stimulus since the (un)awareness can indicate the direction of the contextual effects.

Awareness of an influence can also be described in terms of whether it is covert versus overt (White et al., 2020; see also the *set-reset model* in Martin & Shirk, 2007). When a participant is aware of the potential influence of a first stimulus onto a second one (its influence is overt), then they are more likely to exclude it from their representation of the second stimulus and contrast effects will occur. But if they are unaware of the influence (covert), then they may include it in the representation of the second stimulus and assimilation will be observed (Martin, 1986; Lombardi et al., 1987; Liberman et al., 2007). This raises several questions, notably: how do we determine whether the participant is aware of a stimulus influencing them? If we believe that the Schwarz and Bless (1992) model is generally applicable, then the difference in results between the independently rated stimuli (I) and the single (S) and double (D) rated conditions indicates that the influence of the first stimulus on the second one is overt in some cases, covert in others. But there is no obvious way to check whether an influence is covert or overt other than measuring it directly, which according to our model disturbs the system. One approach is to infer whether a first stimulus is overt versus covert from the direction of the influence of the first one onto the second one, in a post hoc way. For instance, we generally observed contrast effects across our experiments, which could imply that the influence of the first stimulus onto the second is overt. Of course, such a line of reasoning does not offer independent verification of these ideas, but it simply allows a descriptive match between results and theory.

Our results indicated that negative impressions were (generally) more impactful on positive stimuli than positive impressions on negative stimuli. It is therefore important for us to also explore why the strength of these contextual effects differs depending on the valence of the stimuli. In the following paragraphs, we explore several reasons for this.

First, participants may be more attentive to negative stimuli. Early evidence in attentional bias came from Hansen and Hansen (1988) who showed participants an array of happy faces with a single angry face, and an array of angry faces with a single happy face. They found that participants were quicker at selecting an angry face out of a happy array than the contrary. They concluded that attentional processing occurred automatically and that participants were drawn to the angry faces. This finding has been replicated by several researchers (Eastwood et al., 2001; Fox et al., 2000) and by others noting the exception that this only occurred for threatening faces but not for other types of negative faces (e.g., sad; Öhman et al., 2001). Before we move onto other evidence of attentional preference to negative stimuli, it is worth pointing out a limitation with the aforementioned studies. Barrett et al. (2019) challenges the notion of a 'prototypical' facial expression (e.g., simulated expressions). When does someone show a prototypical threatening face? This is likely only under the most extreme situations. Perhaps some of the research outlined above is problematic in that the employed stimuli contained expressions that were both 1) recognisable, but also 2) reflected no aspect of everyday life. Nevertheless, in a series of experiments, Pratto and John (1991) showed that heightened attentiveness to negative stimuli was not limited to just facial processing. They found that participants undertaking variations of the Stroop task exhibit longer latencies when rating negative words compared to positive words. Pratto and John's (1991) findings suggest that events which may negatively affect the individual require greater processing time than events which lead to desirable consequences. Further examples of negative attentional bias have been found in anxious and depressed individuals (Mogg et al., 1995) and in changing fear beliefs in children (Field, 2006).

Second, a negativity bias exists whereby a greater weight is given to negative events, objects, stimuli, and other negative entities. Rozin and Royzman (2001, p.296) offer an intuitive example of a negativity bias, whereby: "… contact with a cockroach will usually render a delicious meal inedible. The inverse phenomenon-rendering a pile of cockroaches on a platter edible by contact with one's favorite food - is unheard of." Indeed, this line of reasoning has been maintained by many psychologists (Baumeister et al., 2001; see the *four aspects of negativity bias* in Rozin & Royzman, 2001; see also *loss aversion* in Tversky & Kahneman, 1991) and provides substantial support for the idea that negative stimuli are more impactful than positive ones. Further evidence is provided by Sheldon et al. (1996) who asked participants to complete a two-week diary containing a battery of well-being measures. They found that participants who reported a 'bad day' also experienced carry over effects. Specifically, they suffered worse days when they reported feeling sick or sad on the previous day. In contrast, participants did not feel better when they reported feeling 'more positive affect or vitality' the day before. Similar results have been found in a variety of settings, such as social evaluations by infants (Hamlin et al., 2010; Hamlin et al., 2007), larger P1amplitudes in event related brain potentials in negative stimuli than positive counterparts (ERPs; Smith et al., 2003), and negative market effects in stock returns (Akhtar et al., 2011). Clearly, there is a persistent negativity bias that has been established in different contexts.

However, there is also evidence that people are generally optimistic and anticipate positive outcomes, even in the face of reasonable negative outcomes (see *unrealistic optimism* in Weinstein, 1980; DeJoy, 1989). This effect is known as the optimism bias (Sharot, 2011). Consider an example of newlyweds estimating their chances of divorce as near-impossibility relative to divorce rates seen across many countries (e.g., ~33% in England and Wales, ONS, 2019; and just over half of all first marriages ending in divorce after 20 years in the US, Copen et al., 2012). This bias refers to the tendency that people overestimate the likelihood of experiencing good events in their life (e.g., a long, happy marriage) whilst also underestimating the likelihood of experiencing bad events (e.g., decline in marital satisfaction, or even divorce; see Lavner et al., 2013). This effect is not only observed when estimating the likelihood of events relevant to oneself, but also when one person compares themselves to another. Specifically, individuals are more likely to anticipate positive outcomes for themselves and think that negative events are more likely for others. In a seminal study, Weinstein (1980) reported that students estimated they were ~44% more likely to own their own home than the other students in the class, ~58% less likely to develop a drinking problem, and 56% less likely to commit suicide than their peers. Other studies have found that drivers consider themselves more skilled than others on the road (Svenson, 1981), students overestimate their future income relative to their peers (Seaward & Kemp, 2000) and individuals who show higher levels of the optimism bias engage in less protective behavioural changes (e.g., social distancing and mask wearing) amid the Covid-19 pandemic (Fragkaki et al., 2021).

A similar effect also exists in that people form beliefs asymmetrically: readily accepting positive information into their beliefs and averting the incorporation of unfavourable information (see the *good news-bad news effect*, Eil & Rao, 2011; Sharot et al., 2012). Sharot et al. (2012, p.1) note that "…people adjust their beliefs regarding their level of intelligence and physical attractiveness when they receive information indicating they are more intelligent and attractive than they had assumed. However, they relatively fail to adjust their beliefs in response to information suggesting they rate lower on these attributes than they had previously thought." Moreover, Sharot et al. sought to identify the specific brain region(s) responsible for this effect. In their study, they presented participants with 40 different adverse life events (e.g., cancer) and asked participants to estimate the probability of this life event occurring to them. Participants were then shown the average probability of that same event occurring to another person living in similar circumstances. Sharot et al. also manipulated this average probability to be higher (bad news) or lower (good news) than a participant's estimation before asking participants to report again their estimations of experiencing the event. The authors were able to selectively eliminate the effect by stimulating a specific brain region (inferior frontal gyrus). This elicited the incorporation of unfavourable information into beliefs of vulnerability in participants (i.e., participants would be more likely to factor in unfavourable probabilities into their own).

It is worth noting that the positive biases discussed above (e.g., good news-bad news effect and optimism bias) are found only when one considers personally relevant information, such as when one anticipates the length of their marriage, future income, or their chances of developing cancer. In contrast, in our experiments, participants generally made stronger negative judgements on an array of personally irrelevant advertisements. This point reconciles our finding that initially negative information seems to have a higher impact on the perception of the second stimulus with the work of Sharot and others regarding the positive biases.

Chapter Four: Temporal Bell

4.1 Introduction

We have devoted considerable time exploring the idea of supercorrelations and the Bell paradigm. The importance of this effort goes hand in hand with the key insight that supercorrelation reveals a kind of association which is stronger than that of perfect classical coordination (perfect correlation or perfect anticorrelation). We have tried to illustrate these ideas in the context of a decision game where a participant was trying to coordinate responses with a hypothetical counterpart. One of the surprising implications of this analysis was that choice statistics could not be modelled by a baseline classical model. By baseline, we mean a model which does not take into account explicitly the differing experimental conditions; and by classical we essentially assume macrorealism, that is, that questions have an answer independently of a measurement. Cognitively, this means that, for example, preferences are concrete, regardless of whether we are asked to articulate them or not. We think this is a promising perspective and provides a fruitful route for further exploring the idea of supercorrelations in the interaction between cognitive agents. But, of course, it is also, as things currently stand, a bit cumbersome: we require two separate questions for the two agents and it seems unlikely that such situations will occur frequently in nature, especially if we also expect the sensitivity to context that is required for supercorrelation.

There is an alternative way to establish supercorrelations and this is what we consider next. It concerns supercorrelations of a single (binary) question but measured at different points in time. Supercorrelations in this case concern the way the values of the questions across different pairs of time points (12, 23, and 13) constrain each other. QT can involve situations which can easily produce such temporal supercorrelations (that is, there are plenty of physical situations which produce temporal supercorrelations). Behaviourally, such supercorrelations can be interpreted on the basis of the underlying quantum models which produce them: they show that decisions at earlier time points change the state and so affect any cognitive processing at later time points. Of course, QT is not the only way to model so-called constructive influences in judgement: QT just offers a particular way to model constructive influences (e.g., see White et al., 2020, for alternative approaches). However, the key point is that the idea of temporal supercorrelations is more general than quantum (in principle, they can be produced by non-quantum systems) and they can reflect changes in the (mental) state as a result of decisions more general than those allowed within QT.

Temporal supercorrelations concern changes in time and so they relate to changes in a state across time. So, first, we consider some evidence that human temporal processing is subject to constructive changes. Note, we are not addressing changes within short periods of time (which might relate more to working memory or dynamical decision making). Rather, we will focus on constructive processes in memory. Second, we will explain the basic ideas concerning temporal supercorrelations and why the basic insight is surprising. Third, we will show a simple quantum system allowing temporal supercorrelations and note that essentially this system is the basis for our initial empirical demonstration. Fourth, we consider in detail a range of complex issues in the study of temporal supercorrelations. Finally, we offer our proposal for how to empirically proceed.

4.1.1 Is Human Memory Constructive?

Fairly briefly, there is considerable evidence that memory is constructive, rather than reflect a noisy, but essentially veridical record of our past experience. We briefly offer a series of insights regarding what can or cannot be constructive, in relation to human memory, and follow with some more detailed discussion. Overall, there is evidence from neuroscience and behavioural studies indicating that memory is constructive from the confluence of current and past experience, in which errors frequently occur, conscious and non-conscious inferences ‘fill in’ details, and in which wholly false memories can (usually non-consciously) arise (e.g., Bernstein & Loftus, 2009; Schacter et al., 2011). Depending on the theorist, the entirety of memory is constructive (because incoming stimulation is interpreted through memories that already exist, making what gets into memory a constructive mix of what was perceived and what is already contained in memory) or the meaning that gets extracted from the experience is constructed although certain "uninterpreted" perceptual details can remain for a short period of time (cf. the distinction between verbatim and gist memory, Reyna, 2008; Reyna & Brainerd, 1998). Constructions have to be generally consistent with what is already in “semantic memory” — as experience is interpreted through this veil of prior knowledge, constructive storage/retrieval processes are constrained by these priors (Howe, 2011; Nelson & Shiffrin, 2010).

We can inquire in more detail what are the properties of constructive memory processes. We have already noted that memory processes ‘fill in details’ (Reyna & Brainerd, 1998) – so, according to existing theory, constructive processes cannot increase or re-introduce uncertainty. False memories confuse similar experiences (e.g., tie colour), create entire events (e.g., UFO’s), but can also be adaptive (e.g., if false memories enhance self-image or prime solutions to complex problems; Howe, 2011). False memories cannot contain things outside an individual’s world view (Howe, 2011). There are several types of false memories (Schacter et al., 2011): Imagination inflation (imagining a novel event can lead to false recollection of an experience, based on similarity between actual and imagined events); gist-based errors (false recall of a novel item, similar to a studied one); post-event misinformation (erroneous information after initial encoding persists). Gist memories can be adaptive but can also enhance false memories for similar items (Schacter et al., 2011). Both positive and negative events can be prone to distortion (Schacter et al., 2011). People appear to recall events (e.g., a story) according to their expectations and pre-existing schemata (Bartlett, 1932).

For example, remembering is very context dependent and this can have some important applied implications. For example, we know that witnesses to a crime often discuss what they have seen with other witnesses (Paterson et al., 2009). Gabbert et al. (2003) investigated memory conformity effects between individuals who witness and then discuss a criminal event. They employed a novel procedure whereby each member of a dyad watches a different video of the same event. Each video was filmed from a different angle and contained unique items that were seen only by one witness. Importantly, in only one of the conditions was the participant able to see the theft of some money. Thus, their protocol was as follows: dyads viewed a film, then discussed the contents and provided a joint account before finally providing an individual account of the film events. A large proportion (71%) of witnesses who had discussed the event went on to mistakenly recall items acquired during the discussion (some claimed that they had also ‘seen’ a non-existent theft despite them not being the witness who were presented with the incriminating video angle). Hope et al. (2008) support these findings, adding that the relationship between the two members of the dyad affects the extent to which they will incorporate misinformation from the other person (varying from strangers to friends to couples). Importantly, dyads involving friends and couples were more susceptible to the memory conformity effect than strangers. Arguably, these effects may be explained by how much one trusts another person, impacting on the extent to which they will accept the incoming information to be true.

Sometimes we retain very vivid and detailed memories of certain drastic world events even though we did not witness the event ourselves (see *flashbulb memories* in Brown & Kulik, 1977). For example, one need only consider where they were when they heard about the 9/11 terror attack in the city of New York. Evidence suggests that an event has to be 'personally relevant' in order for you to form a flashbulb memory (Conway, 1994), but even these memories are subject to distortion (Greenberg, 2004). Greenberg (2004) notes that even U.S president George W. Bush (on at least three occasions) offered inconsistencies on how he heard the news of the 9/11 attacks. For instance, although there was no actual footage of the first plane hitting the world trade centre (at least in the morning of the event), Bush claims there was. Similarly, there are numerous other findings of false details in flashbulb memories ranging from Princess Diana's car crash (44%, Ost et al., 2002), assassinations of Dutch politicians (66%, Jelicic et al., 2006; 66% Smeets et al., 2009), CCTV footage of an explosion of the No. 30 bus in Tavistock Square (7/7 bombings; 39%, Ost et al., 2008), and the assassination of Swedish foreign minister Anna Lindh (64%, Sjödén et al., 2010). Indeed, other emotionally arousing events can lead to a narrowing of perception. Loftus et al. (1987) report the weapon focus effect whereby attentional narrowing occurs when a person is looking down the barrel of a gun (they tend not to remember too many peripheral details).

4.1.2 The Leggett-Garg/ Temporal Bell Inequality

Consider three assumptions: macrorealism, to mean that a system is always at a specific state. For example, consider a question of the guilt or innocence of a suspect in a hypothetical crime. A participant in a psychology experiment receives information across consecutive time points, concerning the guilt or innocence of the suspect. Macrorealism would mean that at each time point the participant must definitely consider the suspect to be either guilty or innocent (if the binary classification of guilt versus innocence appears crude, this is not a problem: we can imagine there is an internal ratings scale for guilt and then macrorealism assumes that the participant is at a specific point on each scale at each time point). The second assumption is NIM. This means that the measurement itself does not disturb the system. So, if we believe in constructive influences in decision making or cognition generally, this assumption appears problematic. The trick is to construct measurement models such that we can separate what one might call ‘principled’ constructive influences from constructive influences which disturb the system because they are too crude in some sense. We want to explore any constructive influences over and above just ‘sledgehammer’ constructive influences. The final assumption is just an arrow of time assumption, that whatever happens earlier can influence what might happen later, but not vice versa.

Next, we define the following quantities:

$\langle Q_i Q_j \rangle = \sum_{n_i,n_j} q(n_i) q(n_j) P(n_i, n_j)$, where $n_i, n_j = \{+,-\}$ ……………………..Equation (12)

This quantity requires some explanation, but it is very straightforward. The straightforward bit is that it can be thought of as a correlation and, really, it is an expectation exactly like the ones we discussed in relation to standard Bell. The n_i etc. indicate the various values the question can take at different times points, denoted by i. So, for example, n_1 is the value the question takes at the first time point. Since we are so far assuming binary observables, these values can be + or -. The quantity $q(n_i)$ is the value we assign to different question outcomes. We just decide that $q(n_i = +) = +1$ and $q(n_i = -) = -1$. Finally, $P(n_i, n_j)$ is the probability that at time points i, j the question will have particular outcomes. For example, we might have that $P(n_1 = +, n_2 = +) = 0.5$, to mean that the probability that the question will have a + outcome at the first and second time point. Note, here we are abusing notation a little bit by using plus, minus signs both for the question outcomes and the values of these question outcomes. In more rigorous treatments, we might say e.g., $n_i \epsilon \{A,B\}$ and $q(n_i) \epsilon \{$+,- $\}$. Invariably, we will devote $\langle Q_i Q_j \rangle$ as C_{ij}, for simplicity and to emphasize the fact that these quantities can be understood as correlations.

As a way to further understand this, suppose we are considering whether to accept a gamble or not. There are different outcomes to the gamble, denoted by *i*, where i=1,2,3,4… (each of these indices denotes one gamble outcome). Let’s say that the payoff for outcome i is *payoff*(*i*) and that the probability for each outcome is *prob*(*i*). We can ask what is the expected value (or expectation) of this gamble. This would be: $\sum_i prob(i) payoff(i)$. The quantity $\langle Q_i Q_j \rangle$ is identical, but for the fact that we are considering joint (indexed by i, j) outcomes. There is an additional step to interpret $\langle Q_i Q_j \rangle$ as a correlation, but this is straightforward (because we have binary observables), if one observes that $\langle Q_i Q_j \rangle$ basically adds the probabilities Q_i, Q_j are the same sign and subtracts the probabilities when they have opposite – which is basically a correlation.

Given the assumptions above, we now state the TB or Leggett-Garg inequality (LGI, Leggett & Garg, 1985) as:

$$C_{ab} + C_{bc} < C_{ac} + 1$$
$$\langle Q_a Q_b \rangle + \langle Q_b Q_c \rangle < \langle Q_a Q_c \rangle + 1$$

where a, b, c denote the three time points of interest. A simplified form if we allow $Q_1 = 1$ (that is, if we assume the initial state to be in a certain position) would be:

$$\langle Q_b \rangle + \langle Q_b Q_c \rangle < \langle Q_c \rangle + 1$$

As in the case of the standard Bell inequalities, the TB version has an almost magical quality that makes us, on first impression, question why it is so special and consider why we should be devoting much thought or effort to study it. The first point to make is that, like the standard Bell, it is derived on the basis of absolutely minimal assumptions. Second, the temporal inequality is a statement of how the correlations across the different time points are bounded: we are told that the correlation across a, b and across b, c has to be bounded by the correlation across a, c plus 1. This bound can be thought of as analogous to the standard Bell bound of 2. It tells us that correlations cannot exceed certain boundaries. $C_{12}+C_{23}<C_{13}+1$ shows that $C_{12}+ C_{23}$ have to be bounded by C_{13} (plus 1), and we can test whether this is the case or not, for any particular system. Simply put, the correlations involving the middle point have to be bounded by the correlations involving the end points.

Let us assume that the assumptions concerning NIM and arrow of time hold. As we shall see, NIM is very challenging to establish (it requires careful methodology), but the arrow of time assumption is trivial. In any case, whether NIM holds or not is essentially a technical/ empirical challenge. So, let us take for granted, for now, NIM and arrow of time, and consider the implications from a violation of the TB inequality concerning macrorealism. If macrorealism holds or does not hold, why would this be interesting or surprising? The answer to this key question is not immediately obvious, but essentially it reduces to this point: A TB violation means it is impossible to have a joint probability distribution for the (assumed possessed) value of the observable across all these time points; it is impossible to concurrently fix the observable values across all time points; there is no classical trajectory, whereby a cognitive observable has specific values across different time points.

The way giving up macrorealism offers a really puzzling view of nature can be illustrated by using an alternative form of the TB inequality, expressed in terms of the number of times an observable changes across specific time points (Atmanspacher & Filk, 2010). For example, $N_{-}(t_1, t_2)$ tells us the number of times the binary question (observable) has changed across the first and second time points, across all possible values that the question can have at these time points. If, for example, at t1 the question is + and at t2 -, then we add one to $N_{-}(t_1, t_2)$. Table 4.1 shows all the possibilities for what $N_{-}(t_1, t_2)$, $N_{-}(t_1, t_3)$, $N_{-}(t_2, t_3)$ can be, depending on the values of the questions at the three time points. It is a straightforward fact of set theory that $N_{-}(t_1, t_3) < N_{-}(t_1, t_2) + N_{-}(t_2, t_3)$ and this inequality is equivalent to the TB one (Atmanspacher & Filk, 2010).

Table 4.1 Values of the binary observable/ question at the three time points, t1, t2, t3 and the corresponding change statistics.

$s(t_1)$	$s(t_2)$	$s(t_3)$	$N\_(t_1,t_3)$	$N\_(t_1,t_2)$	$N\_(t_2,t_3)$
+1	+1	+1			
+1	+1	-1	x		x
+1	-1	+1		x	x
+1	-1	-1	x	x	
-1	+1	+1	x	x	
-1	+1	-1		x	x
-1	-1	+1	x		x
-1	-1	-1			

The violation of the TB inequality, $N_{-}(t_1, t_3) > N_{-}(t_1, t_2) + N_{-}(t_2, t_3)$ means that we may have few changes across t1, t2 time points, few changes across t2, t3 time points, but numerous changes across the t1, t3 time points. Clearly, if the question has fixed (independent of measurement) answers across the three time points this is impossible. So, the TB inequality is essentially about the way a cognitive variable changes across time. If changes are well-behaved, the number of changes across t1, t2 and t2, t3 should be greater than across t1, t3. If the TB inequality is violated, then “the violation of a TB inequality involving a particular observable, at different time points, implies that it is impossible to have a joint probability distribution for the (assumed possessed) value of the observable across all these time points ...it is impossible to concurrently fix the observable values across all time points” (Yearsley & Pothos, 2014, p.7). Put another way, it is as if copies of the observable at different time points are incompatible with each other and so a tabulation of values at different time points is impossible.

The surprising implication for memory is that, for example, recalling a judgement about an observable last week, potentially makes me uncertain about the same judgement the week before, and vice versa, that is, I might become uncertain about judgements which are more recent as well. Of course, the references considered above already point towards a view of memory that is constructive so, in this light, the proposition that querying a memory at a particular time point might impact on the memories (perhaps increasing uncertainty) both before and after might seem less novel? We think not, because existing ideas on constructive influences in memory are specific to the role of these influences in the recollection process: the implications from TB are more mechanistic and relate to the way memories are encoded. We will return to this issue later.

4.1.3 A Simple Quantum System Which Violates The Temporal Bell Inequality

Let us propose that memories can be encoded as quantum states. Why this might be the case is an issue to be considered later. For now, we just assume this proposal as given, with a view to offer a simple demonstration of how QT can violate the TB inequality.

Consider a simple mystery concerning the hypothetical murder of a person, as in the paradigm of Yearsley and Pothos (2016). The trial involves an initial state at which participants assume ignorance regarding the guilt/ innocence of the suspect, some evidence on day 1 and some evidence on day 2. Therefore, the three time points required for a TB setup are {time 1 (a), time 2 (b), time 3 (c)} = {initial state, day 1 evidence, day 2 evidence}. Recall that the TB inequality is $C_{ab} + C_{bc} < C_{ac} + 1$. The assumption of initial ignorance means that the initial state is $\rho = \frac{I}{2}$. Assume one-dimensional observables, for simplicity of illustration (so we can easily generate diagrams as in Figure 4.1). Then,

$$C_{ab} = \Pr(+ + |a, b\,) \cdot 1 \cdot 1 + \Pr(- - |a, b) \cdot (-1) \cdot (-1) + \Pr(+ - |a, b) \cdot 1 \cdot (-1) \\ + \Pr(- + |a, b) \cdot (-1) \cdot 1$$

$\Pr(+ + |ab\,) = \Pr(+ + |initial, Day1\,)$ is the joint probability of having a + (say innocent) initially and a + on Day1 (Figure 4.1a). To compute $\Pr(+ + |ab)$, we have to measure whether innocent or not against the initial state, rotate the resulting state by angle *ab* (this is the change from time 1 to time 2, in the absence of measurement), also measure whether we have innocent or not. Note, in this picture we assume that the question is the same across the three time points and what changes is the state. Note also that this is a counter clockwise rotation and recall that the clockwise rotation operator is $\begin{pmatrix} \cos ab & -\sin ab \\ \sin ab & \cos ab \end{pmatrix}$, so that the counter clockwise rotation operator would be $\begin{pmatrix} \cos ab & \sin ab \\ -\sin ab & \cos ab \end{pmatrix}$. We can, instead, assume that the state is the same but the observable changes – these two pictures are entirely equivalent and preference for one versus the other is just a matter of convenience.

So, we have:

$$\begin{aligned} Prob(++, |ab) &= Tr\left(P_{innocent} \cdot U_{ab}^{\dagger} \cdot P_{innocent} \cdot \rho \cdot P_{innocent} \cdot U_{ab}\right) \\ &= \frac{1}{2} Tr\left(P_{innocent} \cdot U_{ab} \cdot P_{innocent} \cdot U_{ab}^{\dagger}\right) \\ &= \frac{1}{2} Tr\left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \cos ab & \sin ab \\ -\sin ab & \cos ab \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \cos ab & -\sin ab \\ \sin ab & \cos ab \end{pmatrix}\right) \\ &= \frac{1}{2} Tr\left(\begin{pmatrix} \cos ab & \sin ab \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \cos ab & -\sin ab \\ 0 & 0 \end{pmatrix}\right) = \frac{1}{2} Tr\left(\begin{pmatrix} (\cos ab)^2 & \ldots \\ 0 & 0 \end{pmatrix}\right) \\ &= \frac{1}{2} (\cos ab)^2 \end{aligned}$$

Note, we employed here the identity $Tr(|b\rangle\langle a|) = \langle a|b\rangle$, which relates a trace to a dot product.

Note, the same result is obtained if we instead compute $\Pr(++\,|a,b\,) = Tr(P_b \cdot P_a \cdot \rho \cdot P_a) = \frac{1}{2}Tr(|b\rangle\langle b|a\rangle\langle a|) = \frac{1}{2}cos^2(\theta_{ab})$. It may seem curious as to why this is the case, but there is in fact a simple answer (P. Blasiak, personal communication, January 2022). The eigenbases of measurement P_b can be rotated with respect to the first one P_c, i.e., we have $P_b = U_{ab}P_bU_{ab}^{\dagger}$. Thus, using the formula for joint probability in sequential measurement we have:

$$\begin{aligned}\Pr(++\,|c,b\,) &= Tr(P_b \cdot P_c \cdot \rho \cdot P_c) = Tr\left(U_{ab}P_bU_{ab}^{\dagger} \cdot P_c \cdot \rho \cdot P_c\right)\\ &= Tr\left(P_bU_{ab}^{\dagger} \cdot P_c \cdot \rho \cdot P_cU_{ab}\right)\end{aligned}$$

where the latter equation takes advantage of the cyclic property of the trace, e.g., $Tr(ABC) = Tr(BCA)$. So, this seemingly curious property can be seen as passive versus active interpretation of measurements, i.e., instead of measuring at a different angle you can use the same measuring device but rotate the system in the opposite direction (as an aside, this is the same in classical physics). In general, it can be shown (P. Blasiak, personal communication, as above) that the probability to "observe c and then b" is the same as the probability that "the state is positive on Day 1(c) and then also positive on Day 2(b)".

In any case, it is straightforward to compute the remaining probabilities as:

$$\Pr(--\,|a,b\,) = Tr(P_b \cdot P_a \cdot \rho \cdot P_a) = \frac{1}{2}cos^2(\theta_{ab})$$

$$\Pr(+-\,|a,b\,) = \Pr(-+\,|a,b\,) = \frac{1}{2}cos^2(\pi/2 + \theta_{ab}) = \frac{1}{2}sin^2(\theta_{ab})$$

Therefore, $C_{ab} = cos^2(\theta_{ab}) - sin^2(\theta_{ab}) = \cos{(2\theta_{ab})}$

As an aside, using an uniformed initial state might seem unsatisfactory to some readers, since the first rotation is not 'visible' in the diagram in Figure 4.1a. It is straightforward to offer a slightly more complex analysis, such that the we have three time points of evidence, {time 1 (a), time 2 (b), time 3 (c)}= {day 1 evidence, day 2 evidence, day 3 evidence}, and all rotations can be clearly seen. We briefly illustrate this (Figure 4.1b). Suppose now then that the initial state is $|innocent\rangle$. Then, the Day 1 evidence rotates the state in the position indicated by the day 1 ray. Then, $Prob(innocent, day\ 1) = |P_{innocent}|day1\rangle|^2 =$

$||innocent\rangle\langle innocent|day1\rangle|^2 = |\langle innocent|day1\rangle|^2$. But the dot product of two normalised vectors is just the cosine of their angle: $\langle a|b\rangle = \cos\theta_{ab}$, so $Prob(innocent, day\ 1) = \left(\cos\theta_{day1}\right)^2$. Likewise. $Prob(innocent, day\ 2) = \left(\cos\theta_{day2}\right)^2$, $Prob(innocent, day\ 3) = \left(\cos\theta_{day3}\right)^2$. So, initially rotate the state by angle c, as a way to capture the Day 1 evidence, then measure whether innocent or not, then rotate the resulting state by angle cb (this is the change from Day 1 to Day 2, in the absence of measurement), and then measure whether we have innocent or not. This translates to

$$Prob(++,|day\ 1, day2) = |P_{innocent}U_{cb}P_{innocent}|day1\rangle|^2$$

Noting that:

$$|day1\rangle = U_a|innocent\rangle = \begin{pmatrix}\cos a & -\sin a\\ \sin a & \cos a\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}\cos a\\ \sin a\end{pmatrix}$$

(where, as above, we used the rotation matrix in anticlockwise direction), we then have:

$$\begin{aligned}Prob(++,|day\ 1, day2) &= |P_{innocent}U_{cb}P_{innocent}|day1\rangle|^2\\ &= |\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\cos cb & -\sin cb\\ \sin cb & \cos cb\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\cos a\\ \sin a\end{pmatrix}|^2\\ &= |\begin{pmatrix}\cos cb & -\sin cb\\ 0 & 0\end{pmatrix}\begin{pmatrix}\cos a\\ 0\end{pmatrix}|^2 = |\begin{pmatrix}\cos cb & -\sin cb\\ 0 & 0\end{pmatrix}\begin{pmatrix}\cos cb\cos a\\ 0\end{pmatrix}|^2\\ &= (\cos cb\cos a)^2\end{aligned}$$

Overall, assuming a specific state of innocence at time 1 introduces an extra factor in the result. However, the point we want to illustrate here is the expected size of LGI violations; therefore, this elaboration is unnecessary, and we proceed more simply, on the assumption that there are two days of evidence and an initial state of $\rho/2$. That is, $\Pr(+ + |a, b\ ) = \frac{1}{2}cos^2(\theta_{ab})$ etc.

Following standard algebra (P. Blasiak, personal communication), assume an arrangement of rays so that $\theta_{ab} = \theta_{bc} = \theta_{ac/2}$ (Figure 4.1c). Then, $C_{ab} = C_{bc}$ and, $C_{ac} = cos^2(2\theta_{ab}) - sin^2(2\theta_{ab})$. The TB inequality, $C_{ab} + C_{bc} < C_{ac} + 1$, can be rewritten as $C_{ab} + C_{bc} - C_{ac} - 1 < 0$ so there is a violation whenever this quantity is positive. Plotting $C_{ab} + C_{bc} - C_{ac} - 1 = 2C_{ab} - C_{ac} - 1$, with values for C_{ab} and C_{ac} as just above, we see that there are plenty of regions where there is a violation of the TB inequality, even for such a simple system (Figure 4.2). The key result is this: The maximum violation appears to be about 0.5, when the angle is about 0.45. In fact, this result is general for a two-dimensional system, though it is not obvious why this is so from the present analysis.

For general angles, we could rewrite all this as:

$$C_{ab} = cos^2(\theta_{ab}) - sin^2(\theta_{ab}) = \cos{(2\theta_{ab})}$$
$$C_{ac} = \cos{(2\theta_{ac})}$$
$$C_{bc} = \cos{(2\theta_{bc})}$$

Consider the TB inequality again, $C_{ab} + C_{bc} < C_{ac} + 1$, there is a violation when

$C_{ab} + C_{bc} > C_{ac} + 1$ or $C_{ab} + C_{bc} - C_{ac} - 1 > 0$

This can be easily rewritten with the expressions above, as:

$\cos{(2\theta_{ab})} + \cos{(2\theta_{bc})} - \cos{(2\theta_{ac})} - 1 > 0$

noting that:

$\theta_{ab} = Day3 - Day2, \theta_{bc} = Day2 - Day1, \theta_{ac} = Day3 - Day1.$

Figures 4.1a (left), 4.1b (middle), 4.1c (right). The figures illustrate how the mental state vector would change depending on the evidence for guilt presented across the three days of the hypothetical trial.

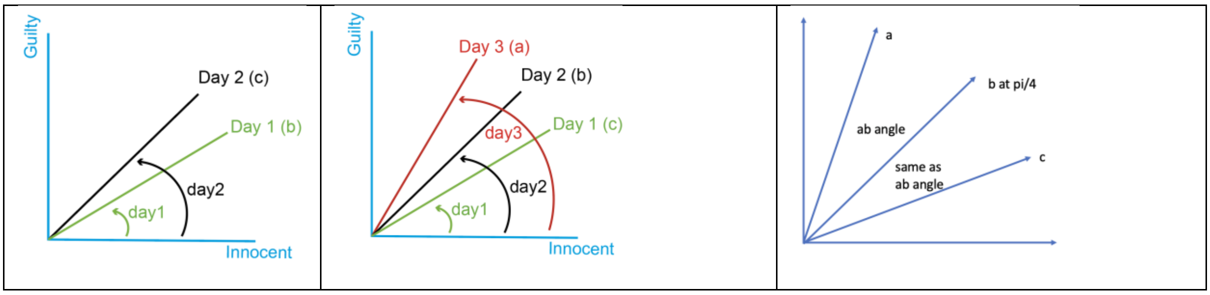
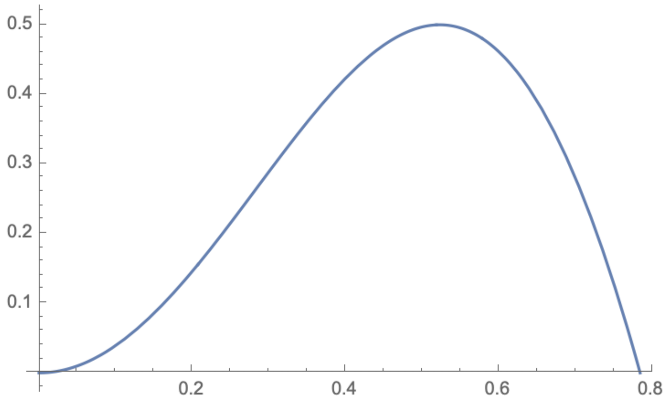


Figure 4.2. The extent of violation of the TB inequality. We are plotting the quantity $2C_{ab} - C_{ac} - 1$ on the vertical axis, while on the horizontal axis we have the size of the angle ab (please see Figure 4.1b).



4.1.4 Broader Considerations

The main conclusion we want to reach so far is that, even for a simple two-dimensional system we can easily identify conditions which violate the TB inequality. The Figures 4.1a, 4.1b considerations can be easily described in terms of requiring small changes across days 1, 2 and days 2, 3, but a large change across days 1,3. This leads to large terms on the left hand side of the TB equation, $C_{ab} + C_{bc} < C_{ac} + 1$, but small terms on the right hand side. We have yet to address why this matters. To do so, we first discuss some relevant ideas from quantum physics and then consider the applicability of these ideas in cognitive theory.

Consider the analysis of Kleinmann et al. (2011). They considered a set of measurements and different products of combinations of measurements. They assumed that there are classical states for each measurement, which perfectly capture different measurement outcomes. If there are a series of classical measurements, what combinations of classical states are needed to reproduce results from a quantum system? This analysis was conducted in relation to the standard Bell inequalities and its point was to examine how much classical resources would be needed to reproduce the statistics from a quantum system.

More presently relevant is a corresponding analysis for a temporal situation. We discuss Budroni et al. (2019). These investigators considered a box with an input, output operation, used at different times. With this setup, a sequence of outputs would be observed. The question is how to reproduce this sequence. If one knows all previous inputs and outputs, one can trivially reproduce any sequence with a simple map: Required Sequence = function (all previous inputs, outputs). However, the key question is whether there more minimal systems which allow one to reproduce particular sequences. Budroni et al. (2019) quantified memory resources as the number of distinguishable states from each probability theory. The point is to identify the minimal system which can produce some correlations. How does a memory-aided model for the system work? In the classical case, one starts with a state and at each step one is free to impose any transformation on the state, allowed by the probability theory and dimensionality (this is what each box does). In the quantum case, one also starts with a state and at each step one can measure on any basis one likes. In either case, at each step one has a new output, given an input. The intention is that by doing this one will end up with the target sequence. To work out the information resources needed from classical versus QT, one then creates an expression S of correlations from the system (which need to be modelled; these correlations are joint probabilities of combinations of outputs, given particular inputs) and then considers how S can be computed in terms of operations possible from each probability system. Specifically, Budroni et al. (2019) considered a two-dimensional classical system, a two-dimensional quantum system, and a two-dimensional general system. The analysis of Budroni et al. (2019) is very complex, but the essential conclusion is that a two-dimensional quantum system can produce stronger correlations (higher values of S) than a two-dimensional classical system. With some care, this conclusion can be taken to mean that if you are interested in a temporal map of some complexity (that is, an association between inputs and outputs, concerning the same question, at different time points), then a quantum approach is more powerful/ expressive, than one based on CPT.

Let us take a step back. Suppose we have the task of designing a memory system for an intelligent agent. Arguably, a very powerful approach would be to employ quantum states. This would mean that at each step, what is encoded is a snapshot of the agent's experience, rather than propositional lists encompassing all possible aspects of experience at each time moment. Why would this be advantageous? Because such a state/ snapshot can then be queried flexibly at future points: in QT, given a state, this can be queried in terms of any possible basis. So, for example, the agent might be focusing on a particular task, but subsequently the sequence of memories/ states could be queried in terms of questions not relevant to the task. The advantage of this immense expressive power has to be balanced with the disadvantage that every time a quantum (or quantum-like) state is queried, it has to change. So, for example, if we ask the agent, ‘how cold were you in the morning’, the response to this question will impact on other recollections concerning this state. Note, a quantum (pure) state has zero entropy, because any state is an eigenstate (that is, it corresponds to a specific response) to *some* question. Recall, a pure quantum state has the form of a vector in a quantum space. A pure quantum state can always be thought of as aligning to the outcome (an eigenstate/ basis vector) to *some* question, even if this question may not have a simple verbal expression.

One might argue that such arguments for quantum advantage (as in Budroni et al., 2019) are irrelevant to considerations concerning intelligent agents (human or otherwise), because for such agents any quantum effects or states are epiphenomenal, they do not reflect real quantum character at the neurophysiological level (Yearsley & Pothos, 2014). However, it is unclear how such an argument can be resolved. Behaviourally, there is a set of processes at the cognitive level, call them A, and there is a set of related processes at the neurophysiological level, call them B. It is possible that capacity restrictions are more stringent for A than B, e.g., if for A there are additional requirements (such as working memory etc.). So, if epiphenomenal quantum-like states offer us some saving at A then this is just as well, regardless of whether any savings are eliminated when the pure states are classically represented at B.

There is an alternative way to make this argument: If I am a ‘quantum’ reasoner (at the epiphenomenal level), I can define conjunction as a sequential projection and then be happy to write Prob(A & then B)> Prob(B) (as in Busemeyer et al.’s, 2011, quantum model of the CF). This will not always be the case, but it will no longer be disallowed. But, if I am a classical reasoner, I can also say Prob(A & B | X) > Prob(B | Y) — I can reproduce the apparent CF, but I have to introduce additional states. This simple example shows that quantum-like representations can offer apparent savings relative to classical ones, even if we completely disregard the issue of implementation at the neurophysiological level.

The bottom line is this: there is some appealing (even if somewhat vague at this point) motivation for the proposal that human memories are snapshots of experience at each point, encoded as quantum states, which can be flexibly queried at future points. Now, of course, this is a very complex proposal. One aspect of this proposal is that encoding of experience does involve quantum states. The corresponding test is exactly the test of the TB inequality: if we have quantum (by which we really mean quantum-like) states and processes, then it would be possible to violate the TB inequality. Violations of the TB inequality would reveal temporal correlations stronger than what could be obtained from classical systems (cf. Budroni et al., 2019). Again, this has relevance to cognitive modelling and architecture: a 'good' memory system is arguably one which maximizes information transfer across time points. A violation of TB inequality can be interpreted as showing increased information transfer, at least relative to a system which is constrained by macrorealism.

We summarise this discussion by presenting the key questions which are guiding us: Are constructive influences in temporal judgements of the specific kind predicted by QT? Can we challenge a simple notion of logical progression in memory, so that measurement at a point creates interference both before and after? Is memory encoded as quantum-like states, corresponding to snapshots at different times? It is hugely implausible that we store information about everything at all times. It would seem more efficient that we have a quantum state, which can be queried flexibly depending on what we want at later times.

4.1.5 Empirical Challenges

The essential objective of a TB inequality test is to test for quantum character in a system, since, recall, the main assumption which we wish to challenge with an observation of a TB violation is that of macrorealism (if the macrorealism assumption has to be aborted, then this increases our confidence that the states have quantum-like character). However, if we observe a violation of the TB inequality, we need to ensure that the measurement is not disturbing.

Let us consider this point again: if we observe a violation of the TB inequality, then this means that the measurements changed the system. This could be because of one of two reasons. The first and more interesting reason is because we cannot uphold the assumption of macrorealism, the states have quantum-like character, and a measurement changes the system. The second and less interesting reason is because the measurement is 'disturbing', that is, it is like a sledgehammer which, when performed, changes the system. In physical terms, a disturbing measurement might be something like say a device which measures the momentum of an object, by capturing it in a trap of some sort: the measurement might offer a high degree of precision, but in capturing the object, it changes its momentum post-measurement.

In order to test for whether a measurement is disturbing or not, we have to employ the so-called non-signalling in time (NSIT) tests. For example, if we have three time points across which we measure our system on a binary question, let us first define:

$$\delta(n_3 = +) = P(n_3 = +) - \sum_{n_2} P(n_3 = +, n_2)$$

$$\delta(n_3 = -) = P(n_3 = -) - \sum_{n_2} P(n_3 = +, n_2)$$

$$\delta(n_2 = +) = P(n_2 = +) - \sum_{n_1} P(n_2 = +, n_1)$$

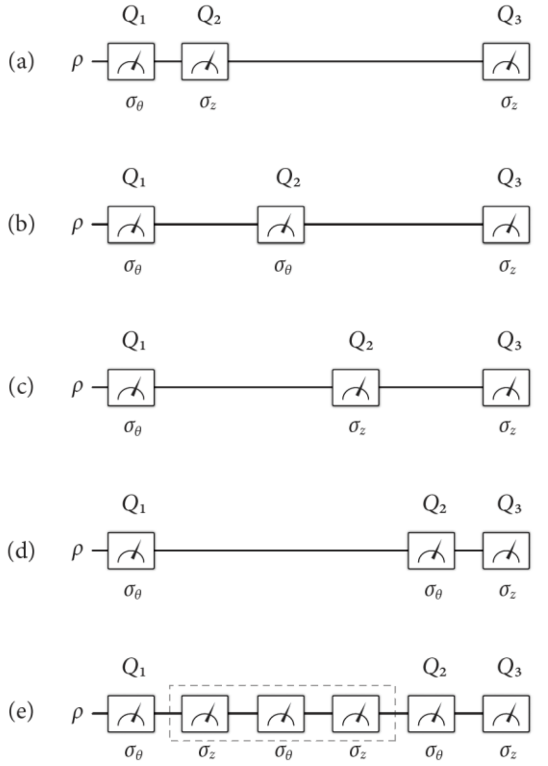
$$\delta(n_2 = -) = P(n_2 = -) - \sum_{n_1} P(n_2 = +, n_1)$$

The NSIT requirement is that all these quantities are zero. That is, the next measurement is the same, whether we measure at the previous time point or not. Note, the marginals (e.g., $P(n_3 = +)$) are measured in experiments for which there are no prior measurements. The conjunctions (e.g., $P(n_3 = +, n_2 = +)$) are measured in experiments for which there was a measurement in the previous time point. Observing that the delta quantities are zero just means that the previous measurement does not affect the state.

But, if our theory is that the state has quantum-like structure, how is it possible for the previous measurement to not affect the system? This is a highly non-trivial problem and essentially the reason why tests of TB inequalities have been less forthcoming than corresponding tests of the standard Bell inequalities. We consider two approaches to the problem from physics, examine how we could adapt these approaches to behavioural sciences, and finally offer a proposal which is unique to behavioural sciences. Both physics approaches are focussed on how we can avoid disturbance from quantum processes; in all cases, it is assumed that additional care must be taken to ensure that the measurement is ‘gentle’ enough so that there is no disturbance from the act of measurement, regardless of quantum character versus not. Note, physicists have considered all sorts of exotic solutions to the problem of NSIT. The options we discuss below are the ones which we consider more transferable to behavioural experimental protocols. As examples of other ideas, which we believe are less presently relevant, we note the proposal that instead of directly measuring the system of interest we measure an ‘ancilla’ system, which is a system coupled with the one of interest (e.g., Emary et al., 2015; Halliwell, 2016). But such an approach appears implausible behaviourally. Physicists have even considered the notion of negative probabilities, but there is debate in terms of whether corresponding measurements can circumvent disturbance from quantum processes (Emary et al., 2015).

Wilde and Mizel (2012) proposed that we can avoid the problem of disturbance from quantum processes, if we ensure that measurements occur at these time points such that the state is an eigenstate of the observable we are measuring. Looking at Figure 4.3 (from their paper), it is assumed that at different points a different question is asked – that is, in their setup the state is the same and what evolves is the question (this is entirely equivalent to assuming that the question is static with time and what evolves with time is the state). In Figure 4.3, different rows correspond to different experimental conditions. For example, row (a) means that we make the three measurements as shown etc. The key feature in Figure 4.3 is that within each row there are always two successive measurements such that the second one will ‘receive’ an eigenstate from the previous one. For example, consider row (a). The first measurement of σ_θ produces an eigenstate of that question. The second measurement of σ_z produces an eigenstate of σ_z. But, importantly, the final measurement is also of σ_z, so that, as far as quantum processes are concerned, this third (final) measurement should be non-disturbing relative to the previous one. Because this protocol always involves a measurement which ‘sees’ an eigenstate (produced by a previous measurement), it is possible to devise a schedule of measurements that fulfils the NSIT condition. This is an elegant proposal, for physicists at least, but it relies on an assumption that it is possible (even if theoretically) to completely eliminate disturbance (quantum or otherwise) from measurements. Behaviourally, this might be a tricky requirement.

Figure 4.3. Protocol for avoiding NSIT, as seen in Wilde and Mizel (2012).



We follow the exposition from Emary (2017). Let us assume that measurements are invasive. Let us assume that the question outcome at time 1 is just 1, for a binary question Q. This is a straightforward simplifying assumption, concerning the initial. Also, make the following definitions (similar to the ones above, but we repeat them here, using the $Q_1 = 1$ assumption and the notation from Emary, 2017; note the subscript in Q indicates the time point, so $Q_1 = 1$ just means that the question outcome at time 1 is 1).

$$\langle Q_i \rangle = \sum_{n_i} q(n_i)P(n_i)$$

This expectation value concerns the correlations between times 1 and 2 and 1 and 3; for these two pairs of times, by assumption $Q_1 = 1$, and so the expectation values can be written more simply. Specifically,

$$\langle Q_1 Q_2 \rangle = \langle Q_2 \rangle = \sum_{n_2=A,B,C} q(n_2)P(n_2)$$

$$\langle Q_1 Q_3 \rangle = \langle Q_3 \rangle$$

Here, we have jumped ahead a little bit and snuck in the assumption that the possible states at each time point are three (A, B, C), not two as in the standard TB set up. The reason for this would become apparent shortly. We also have:

$$\langle Q_3 Q_2 \rangle = \sum_{n_3, n_2} q(n_3) q(n_2) P(n_3, n_2)$$

Note again that $n_3, n_2 = \{A, B, C\}$. Again, A, B, C are the possible states which can be measured at time 3 or time 2.

Next, define the quantity:

$$K = \langle Q_2 \rangle + \langle Q_3 Q_2 \rangle - \langle Q_3 \rangle$$

Recall the form of the TB inequality that we encountered above, $C_{ab} + C_{bc} \leq C_{ac} + 1$. So the expression for K just reflects a rearrangement of the correlators (as the expectations C_{ab} etc. are called). We can simply just rewrite the TB inequality as

$$K \leq 1$$

So far, we have said nothing new. Now, let us consider again the signalling quantifiers and a related quantity concerning their sum:

$$\delta(n_3) = P(n_3) - \sum_{n_2 = A,B,C} P(n_3, n_2)$$

$$\Delta = \sum_{n_3 = A,B,C} |\delta(n_3)|$$

With $\Delta = 0$, we satisfy the NSIT condition; that is, with $\Delta = 0$, there is no disturbance. Suppose however that there is some disturbance. Emary (2017) showed how we can rewrite the Temporal Inequality condition, taking into account a Δ which is not 0:

$$K \leq 1 + \Delta$$

So, essentially, the signalling quantifiers work like a correction in the TB inequality. Unfortunately, this corrected TB inequality can never be violated! This is a key result which Emary (2017) showed. As Emary (personal communication, 2020) put it:

"Say you observe an LGI violation. You then would like to rule out the clumsiness of your measurements. One way of quantifying the clumsiness of your measurements is via the NSIT/Quantum witness quantities. And what you see is that the degree you violate the LGI can be explained [by] the degree to which you violate the NSIT equalities. Thus, your LGI violations can be explained by the clumsiness shown in the NSIT experiments. Or, …If you modify the LGI to include the clumsiness observed in NSIT experiments, these modified ones can never be violated."

This is a key restriction which casts doubt on the extent to which it is meaningful to study violations of the TB inequality in behavioural settings, at all, that is, on the assumption that some disturbance will always be present. Emary's (2017) analysis so far concerns a standard set up of unambiguous measurements, that is, each measurement reveals a specific state for the system. Emary (2017) proposed that it is possible to get round the above problem by using ambiguous measurements (e.g., A or B – more details shortly). That is, with ambiguous measurements, it is possible to fulfil NSIT, satisfy another condition (called ESIT – equal signalling in time—which means that the disturbance from ambiguous and unambiguous measurements is the same), and violate the TB inequality in this form $K \leq 1 + \Delta$.

Why might we be able to violate the disturbance-corrected TB inequality with ambiguous measurements, but not unambiguous ones? Because in the former case the disturbance to the system is less than in the latter case and so, in the former case, we can fix the dynamics in a way that we avoid any disturbance due to quantum processes (quantum measurement).

We describe the building blocks of his analysis one by one. The ambiguous case TB inequality is given by:

$$K_A \leq 1 + \Delta_A$$

The ‘A’ denotes ambiguous and it will come up on most relevant quantities in what follows.

The ambiguous measurement only occurs at the second time point; the measurements at the first and third time points are still unambiguous as before.

The signalling quantifiers are given by:

$$\Delta_A = \sum_{n_3} |\delta_A(n_3)|$$

$$\delta_A(n_3) = P(n_3) - \sum_{n_2} \mathcal{P}(n_3, n_2)$$

The main difference with the unambiguous case is this: in the unambiguous case, all probabilities are actual probabilities. That is, e.g., when we write $P(n_3)$, this is the probability that we have a particular state at the third time point. In the ambiguous case, we can also talk about inferred probabilities, which are the probabilities about different states that we infer from our ambiguous measurements. That is, $\mathcal{P}(n_3, n_2)$ are inferred probabilities. For example:

$$\mathcal{P}(A) = \frac{1}{2}P(A \cup B) + \frac{1}{2}P(A \cup C) - \frac{1}{2}P(B \cup C)$$

Each disjunction is an ambiguous measurement and we would have these at time 2. For example, $P(A \cup B)$ would be computed from measuring whether *A* occurs or *B* occurs (this would be an example of a question in the paradigm). The logic of the expression for the ambiguous measurement is simply that we take into account the probabilities for *A* from the ambiguous measurements involving *A* or *B* and *A* or *C*, but subtract the probabilities for B or C. As another example:

$$\mathcal{P}(n_3, A) = \frac{1}{2}P(n_3, A \cup B) + \frac{1}{2}P(n_3, A \cup C) - \frac{1}{2}P(n_3, B \cup C)$$

This is the joint probability of a particular state at t_3 and particular ambiguous measurements at t_2.

There are two issues to consider next. First, the distinction between ambiguous and unambiguous measurements is not straightforward. It will become clearer if we discuss the necessary experimental protocol. Second, how does QT achieve NSIT, ESIT, and a TB violation? We will outline the predictions from Emary (2017), which could be the basis for modelling work in cognitive settings too.

As noted, we first consider the necessary empirical set up (see also Wang et al., 2018). In order to have ambiguous measurements, we need at least three unambiguous states, and this is the approach we adopt. The three unambiguous states can be denoted as A, B, C. Each state is associated with an outcome of the binary question of interest. We set $q(A) = -q(B) = q(C) = 1$. In an empirical situation concerning a hypothetical crime, the three states could concern information about three distinct situations, two of which would indicate (say) innocence and one of which would indicate guilt. In all cases, we set up the system initially so that $Q_1 = 1$ (which means that the value of the binary observable at time 1 is 1). So, we do not need to measure at time 1.

In the unambiguous case, all we then need to do is measure for state A, B, or C at time 2 (this would give us all the possible values of n_2; e.g., 'Is A true?') and for each of these possibilities measure for state A, B, C at time 3. For example, $Prob(n_2 = A)$ is the proportion of yes answers to the question 'Is A true?'; also, $Prob(n_2 = A \,\&\, n_3 = A)$ is the proportion of cases for which we have a yes answer to both questions. These measurements give us nine different conditions. Additionally, we need three more conditions, for the signalling quantifiers, such that we only measure at time 3. Note, it is essential for such TB experiments to have several, separate conditions, each time concerning two times to measure. If we measure at all three time points within a single condition, then the TB inequality will just be satisfied (since in such a case there is macrorealism by default: at all three time points, there would be a specific value for the observable).

In the ambiguous case, at time 2, instead of unambiguous measurements, we now have ambiguous ones. If there are three overall possibilities for the state (A, B, C), then the possibilities for the unambiguous measurement are three as well A or B, A or C, B or C. Therefore, we also have nine conditions in the ambiguous case as well and measurement of e.g. $Prob((n_2 = A\ or\ B)\ \&\ n_3 = A)$ is the proportion of yes answers in the corresponding condition. Note, we do not need anything else for the signalling quantifiers, since the three conditions for which we just measure A, B, or C at time 3 are identical to those just above (for unambiguous measurements). With this protocol, we can measure all the quantities needed to compute NSIT and then test for violations of the TB inequality.

There is a question of how it can be that, in QT, ambiguous measurements can support the concurrent fulfilment of both NSIT and allow for violations of the TB inequality. We present the analysis of Emary (2017), without attempting to explain in detail his various equations – rather, we just state them, as this would be needed to implement his approach computationally; that is, given information about the dynamical change between times 1 and 2 and (separately; they don't have to be identical) between times 2 and 3. Closely following the empirical requirements, we assume to have three possible unambiguous states A, B, C. $q(A) = -q(B) = q(C) = 1$, and further assume $Q_1 = 1$. That is the answer to the binary question at time 1 is 1. Initialize the system in state so:

$$\rho_1 = |C\rangle\langle C| = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Given state ρ_i at state at time t_i, state at time t_j is given by

$$\rho_j = \Omega_{ji}[\rho_i] = U_{ji}\rho_i U_{ji}^{\dagger}$$

where:

$$U = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{pmatrix} \cdot \begin{pmatrix} \cos\chi & 0 & \sin\chi \\ 0 & 1 & 0 \\ -\sin\chi & 0 & \cos\chi \end{pmatrix} \cdot \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

This is a simple expression of dynamical evolution of the state in a three-dimensional space. Specifically, note there are free parameters for rotation of the state along any of the three possible axes.

In the unambiguous case, measurement is projective, which is essentially the assumption that post-measurement we are certain of what the state is. We posit projectors Π_n, e.g., $\Pi_1 = |1\rangle\langle 1|$, where $|1\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. Then, we have is (allowing for violations of NSIT):

$$K^{QM} = \sum_{n_3,n_2} [q(n_2) + q(n_2)q(n_3) - q(n_3)]P(n_3, n_2) - \sum_{n_3} q(n_3)\delta^{QM}(n_3)$$

The signaling quantifiers can be given with a very simple expression:

$$\delta^{QM}(n_3) = P(n_3) - \sum_{n_2} P(n_3, n_2) = \sum_{n,n'\neq n} X(n_3, n, n')$$

Note, the summation for X involves AB, AC, BC, BA, CA, CB. And the disturbance correction is:

$$\Delta = \sum_{n_3} |\delta^{QM}(n_3)|$$

For K^{QM}, summation is over all possible state combinations at n$_2$, n$_3$, e.g., A2A3, A2B3, A2C3, B2A3…(the notation here means that, for example, we have state A at time 2 and state A at time 3). Joint probabilities are computed as (noting that $\Pi_n = \sqrt{\Pi_n}$):

$$P(n_3, n_2) = Tr\{\Pi_{n_3}\Omega_{32}[\Pi_{n_2}\rho_2\Pi_{n_2}]\}$$
$$\rho_2 = \Omega_{21}[\rho_1], P(n_3) = Tr\{\Pi_{n_3}\Omega_{32}[\rho_2]\}$$

For $\delta^{QM}(n_3)$ and other expressions, we also need:

$$X(n_3, n, n') = Tr\{\Pi_{n_3}\Omega_{32}[\Pi_n\rho_2\Pi_{n'}]\}$$

Recall, the disturbance-corrected TB inequality is: $K^{QM} \leq 1 + \Delta^{QM}$. And we have a violation when $K^{QM} - 1 - \Delta^{QM} \geq 0$. Emary's (2017) main point is that this is impossible (either for the quantum prediction or in general).

For the ambiguous case, we only need to adapt the measurement at time point 2. We are measuring for three possible outcomes $A \cup B$, $B \cup C$, $A \cup C$. Instead of projectors, for a particular ambiguous state a, we have POVMs, $F_a = \sum_{n_2} c_{an_2} \Pi_{n_2}$. For example, $F_{A \cup B} = 0.5\Pi_A + 0.5\Pi_B$ etc. and $Prob(A \cup B) = Tr(F_{A \cup B} \rho_2)$. Note $F_a = M_a^2$, $M_a = M_a^\dagger = \sum_n \sqrt{c_{an}} \Pi_n$. For example, $M_{A \cup B} = \sqrt{0.5}\Pi_A + \sqrt{0.5}\Pi_B$. That is, we can see how the constructs corresponding to these ambiguous measurements boil down to combinations of projectors. The correlator is (note the A subscript indicates ambiguous):

$$K_A^{QM} = \sum_{n_3, n_2} [q(n_2) + q(n_2)q(n_3) - q(n_3)]P(n_3, n_2) - \sum_{n_3} q(n_3)\delta_A^{QM}(n_3)$$
$$+ \sum_{n_3, n_2} [q(n_2) + q(n_2)q(n_3) - q(n_3)]\kappa(n_3, n_2)$$

Note all the summations in the above concern *unambiguous* states. Emary (2020) offers a particular instantiation of the ambiguous measurement setup in his Section V and we follow this scheme here. The main issue is how to compute the signalling quantifiers (see just below) and the quantity $\kappa(n_3, n_2)$ in the above expression. This is given by:

$$\kappa(n_3, n_2) = -\delta_A^{QM}(n_3) + \sum_{n \neq n_2} [X(n_3, n_2, n) + X(n_3, n, n_2)]$$

In this equation, $X(n_3, n, n')$ is the same as in the unambiguous case. Note also that there is a discrepancy here between Equation (23) in Emary (2017) and the equation just above: in the paper, the summation extends across all possible states (A, B, C), whereas in the equation just above we exclude state n_2. We have confirmed with Emary that this was a minor typo in the paper. But what are the signalling quantifiers? Again, with the particular realization in Emary's (2017) Section V, we have:

$$\delta_A^{QM}(n_3) = \frac{1}{2}\delta^{QM}(n_3)$$

This allows us to compute the disturbance correction as:

$$\Delta_A^{QM} = \sum_{n_3} \left|\delta_A^{QM}(n_3)\right|$$

Then, the disturbance corrected TB inequality in the ambiguous case is $K_A^{QM} \leq 1 + \Delta_A^{QM}$ and we have a violation when $K_A^{QM} - 1 - \Delta_A^{QM} \geq 0$.

This model varies predictions depending how the dynamics between times 1, 2 and times 2, 3 occur. The way Emary (2017) approached parameter search was by first identifying the angles which guaranteed $\Delta_A^{QM} = 0$ and then, given this more restricted range of angles, identify the angles which allow for the max violation of the TB Inequality. This is achieved (see also Wang et al., 2018) when the dynamics from t_1 to t_2: $\theta_1 = 0.831\pi$, $\chi_1 = \chi_2 = 0.688\pi$, $\phi_1 = \phi_2 = 0.423\pi$; and the dynamics from t_2 to t_3: χ_2, ϕ_2 as just above. With these angles, for $0.677\pi \leq \theta_2 \leq 0.983\pi$, $K_A \geq 1 + \Delta_A$, and thus we find violations of the modified ambiguously measured TB inequality. Prediction from the quantum model is that max violation at $\theta_2 = 0.831\pi$, for which $K_A^{QM} = 1.464$. Note, the maximum violation for a quantum, unambiguously measured, two-dimensional system, that is the maximum value of $K^{QM} - 1$ is 0.5, a little higher than 0.464, but of course this value does not take into account disturbance. Note, this configuration of angles also confirms the ESIT condition (which, recall, is the idea that ambiguous and unambiguous measurements are equally disturbing). We have not considered the ESIT condition so much in the above discussion partly because, in Emary's (2017) analysis, it is subsumed by some of the other conditions.

Finally, behaviourally, there is a possibility for how to avoid the complexities from the NSIT condition altogether. Recall that the main problem with NSIT is that a prior measurement may have disturbed the system, so that any subsequent measurement is confounded. That is, a measurement at time 2, for example, may not be the case with and without a measurement at time 1. In QT, this is a particularly distinct possibility, since of course measurements collapse the state – and so disturb the system. However, behaviourally, we could simply ask participants for how a particular observable has changed across two time points. For example, in the case of the hypothetical murder case, we could ask them a question like, how much would you say that the guilt of a person has changed between the information you received on day 1 and the information you received on day 2. Why would we expect such a change measurement to be equivalent to expectation values?

This can be explained in the following way. Note first the straightforward point that in a usual TB setup, we would be counting {+, -} occurrences and infer probabilities from these. In the behavioural paradigm, participants do not provide probabilities even, but ratings for {+,-} verdicts. However, ratings can be converted to probabilities via a linear transform (there is absolutely no need to complicate things via assuming non-linear functions, at least in the first instance of this analysis). Given this picture of equating participant ratings with probabilities, then we can ask how the change judgements (interpreted now as changes in guilty/ innocent verdicts) correspond to correlations in the LGI.

Let us introduce some simple notation, namely that the suspect on each day is either innocent, $v_d = +1$, or guilty, $v_d = -1$, where 'v' stands for 'value' and 'd' stands for 'day'. The correlations between two days x, y are then computed as

$$C_{xy} = \sum_{v_x, v_y} v_x v_y P(v_x, v_y) = \sum_{v_x = v_y} v_x v_y P(v_x, v_y) + \sum_{v_x \neq v_y} v_x v_y P(v_x, v_y)$$

When $v_x = v_y$, $v_x v_y = +1$ and when $v_x \neq v_y$, $v_x v_y = -1$. So, we have

$$C_{xy} = \sum_{v_x = v_y} P(v_x, v_y) - \sum_{v_x \neq v_y} P(v_x, v_y)$$

So far, we have just expressed the correlation (expectation value) concerning the verdicts across the two days, x, y.

We are next interested in the change from day x to day y, which we can quantify as $\Delta = +1$ for change and $\Delta = -1$ for no change. Then, the expectation for change can be written as

$$\langle \Delta \rangle = -1 \sum_{v_x = v_y} P(v_x, v_y) + 1 \sum_{v_x \neq v_y} P(v_x, v_y)$$

It follows that $C_{xy} = -\langle \Delta \rangle$. Crucially note that the key issue is just whether there is a change or not, not the direction of the change.

If change is encoded instead so that $\Delta' = +1$ for change and $\Delta' = 0$ for no change, we simply have $\Delta = 2\Delta' - 1$ and so $C_{xy} = -2\langle\Delta'\rangle + 1$.

Note a slightly curious aspect of this formulation, namely that this picture does not distinguish between whether the change is one of assuming initially the suspect is guilty and changing to assume he is less guilty or assuming initially the suspect is innocent and changing to assume he is less innocent. That is, consider asking participants:
$\Delta'' = +1$ – the suspect is judged innocent compared to a guilty verdict before.
$\Delta'' = -1$ – the suspect is judged guilty compared to an innocent verdict before before. But the distinction between change towards guilt versus change towards innocence does not matter, because the C_{xy} values are sensitive only to whether there is change versus not. Therefore, Δ'' can just be recast onto a variable analogous to Δ' and the above formula for conversion to C_{xy} employed.

In the first empirical demonstration we adopt this approach. Recalling that:

$$C_{ab} + C_{bc} < C_{ac} + 1$$

In order to violate the inequality, we need little change across C_{ab} and C_{bc}, but large change across C_{ac} (so that this correlation is low and cannot bound the correlations across the immediately sequential time points). For example, in a hypothetical murder mystery, we can start with a suspect being innocent (time point *a*), then offer weak evidence that he is guilty at time point *b* and weak evidence that he is guilty also at time point *c*. It is possible that between start and c the accumulated change would be perceived as more strong, than the sum between start and *b*, and *b* and *c*. Note, in physics such experiments are carried to a scale and precision that makes it less necessary to employ statistical methods for establishing reliability. However, in psychology, partly because of limited sample sizes, statistical methods are necessary. The correlations in the TB inequality are computed from counts between participants, so there is no straightforward way to employ the inequality above directly. However, it can be shown that this inequality is exactly equivalent to:

$$N^{-}(t_a, t_c) \leq N^{-}(t_a, t_b) + N^{-}(t_b, t_c)$$

where N^- counts the number of changes in the binary observable that is being measured, across the corresponding time points (e.g., Atmanspacher & Filk, 2010). This inequality can be examined with a chi-squared test, comparing the total counts for $N^-(t_a, t_c)$ versus the total counts for $N^-(t_a, t_b) + N^-(t_b, t_c)$, across all participants. Lack of equality (in the right direction) would indicate a violation of the TB inequality.

The final issue to consider is that so far we have focused on this form of the LGI, $C_{12} + C_{23} \leq C_{13} + 1$, but there are a few equivalent relations, notably, $C_{12} + C_{13} \leq C_{23} + 1$, $C_{13} + C_{23} \leq C_{12} + 1$, $-C_{12} - C_{23} - C_{13} + \leq 1$ (Halliwell, 2014). While an LGI violation would have equivalent implications regardless of the specific form of the LGI, the form of the inequality best matching the empirical paradigm we will shortly present is $C_{12} + C_{23} \leq C_{13} + 1$. So for now we will focus on this form.

4.1.6 Interim Summary

Conceptually, this is the most involved topic we will address in this thesis. We will now summarize some of the main ideas. If one takes into account clumsiness (disturbance) in measurements, then any TB violations can be explained by the clumsiness shown in the NSIT experiments. That is, it is not possible to observe TB violation independent of clumsiness. If one modifies the TB inequalities to include the clumsiness observed in NSIT experiments, these modified ones can never be violated. As long as one is even a little bit clumsy, then one can never break the TB inequalities to a degree higher than what you would have just from NSIT violations (Emary, 2020). Wilde and Mizel (2012) also address this problem and their solution is to set up the measurement regime so that in all instances there are back-to-back identical measurements. So, if the measurement is disturbing in the quantum constructive sense, then there would be no effect (because the subsequent measurement would 'see' an eigenstate from the prior one). But if the measurement is disturbing in some other way, then of course such a regime would not protect against this.

One question is how the TB inequalities tell us something different from the standard Bell inequalities. "The underlying assumption behind the bounds of both the Leggett–Garg and Bell inequalities is the existence, independent of measurement, of a joint probability distribution that can provide information on all relevant marginals…" (Emary et al., 2015, p.9). Of course, in standard Bell, one has two systems, which could be called Alice and Bob, and a tensor product space. In TB, one has a single system, a qubit which evolves in time. In standard Bell, one measures different observables in each system. In TB, one measures the same observable at different times (what changes is the evolving system).

So, if we can observe a violation of the TB inequality, how should this be interpreted? Well, the interpretation starts essentially from the converse of macrorealism: with a violation, then we no longer assume that the question has a specific value at each time point, that is, there is no classical trajectory of the observable across time. Note, a classical trajectory is the most straightforward intuition for how memory might be organised, whereby a question has specific values across different time points (Atmanspacher & Filk, 2010; Yearsley & Pothos, 2014). Regarding memory, unless memory is specifically probed for an observable at a specific time, the corresponding memory does not exist at all. This of course also implies that memory is essentially constructive, but in the very specific sense of how measurements can change a quantum system. Probing a memory might create interference with other memories prior to it. This is perhaps unsurprising, given existing ideas. However, the quantum-like picture additional requires that probing a memory might create interference with more recent memories too.

The lack of precise values for memories may seem counterintuitive. Why would a memory system set up so that probing any memory might lead to interference for more recent and older memories? A classical state stores two bits. Violation of the TB inequality supports the possibility that, instead, memory is organised in quantum-like states. For a binary question, we would therefore have qubits. The information in a qubit is far more than that in a bit. The exact comparison is complicated by various considerations, but the essential idea is this: for bit, there are two discrete possibilities (up or down). For a qubit, in the two-dimensional space where it is represented, we can have an infinite number of possibilities intermediate to the two eigenstates in a chosen basis. This is part of the answer why we think it makes sense for memory to be encoded in quantum-like states. Additionally, various results show that a violation of TB inequality implies higher correlations between time points than would be allowed classically (cf. Budroni et al., 2019). So, if you are trying to maximize information transfer between the present time and a previous time (that is, maximize memory recall), then, again, it appears that quantum like states may offer an advantage.

Finally, let us briefly take a step back and ask ourselves: do these experiments really concern memory One would be right to point out that it is not straightforward to interpret the results in terms of implications for memory as such, as opposed to, for example, decision making. A more accurate way to understand the paradigm might be concerning the way decisions are driven from memory representations, since we look at how the evidence at the three different time points in the experiment impact on decisions. The link with memory is still in terms of how these memory representations might be affected by recollection processes – such recollection processes are necessary for the decisions required in the experiment (or at least, this is what we assume, if participants perform the task as intended). In turn, recollection processes test the key prediction from QT, if the relevant memory representations are quantum-like in nature: if this is the case, then recollection should change the representations, and the TB inequalities should be violated. Put in slightly different terms, constructive memory processes would indicate that 'macrorealism' is violated, that is, the assumption that a system is always at a specific state cannot be retained. Overall, while we would like to highlight that the link to memory is somewhat indirect, we hope that readers will agree that the implications from our results do bear on how we understand memory processes (specifically in terms of a particular way in which constructive memory processes arise). We stress that these considerations are still speculative, even if supported by the relevant work (in physics) so far.

4.2 Pilot

We first conducted a pilot study to determine the strength of different pieces of evidence in the hypothetical modern mystery in our experiments. We used pieces of evidence from Yearsley and Pothos (2016) and introduced additional statements that indicated guilt and innocence for the hypothetical suspect. Note, the evidence comprised of guilty and innocent versions and was specifically designed to capture situations in which the suspect could be deemed innocent or guilty (e.g., fingerprints on an item may implicate a suspect or not; see Materials below). The strength ratings also provided a useful benchmark for future experiments by indicating the size and direction of participants' evaluation.

4.2.1 Method

4.2.1.1 Participants

Participants were recruited using Prolific and we restricted sampling to UK nationals. They were paid £1.50 for their contribution. Sample size was set a priori to 200 participants; but, after screening our data for partial completions and failed attention checks, this was reduced to 198 (97 males, 96 females and 5 others who self-identified as 'other'). Participants were between 18 and 56 years old (M_{Age} = 25.52 years old, SD = 7.29). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with most participants reporting 4 or 5 (*n* = 186) and others (*n* = 12) reporting 3 or lower.

4.2.1.2 Materials

4.2.1.2.1 Evidence

Forty pieces of evidence were selected to check the strength of each statement, thirty of which were adapted from Yearsley and Pothos (2016) and the other ten were new for this pilot. There were twenty original statements, but we devised alternative pieces of evidence to indicate an opposing verdict. For example, a statement that might indicate the suspect's guilt, such as, "[the victim] had arranged a number of social engagements for the week after his death" was revised to indicate the suspect's innocence, "[the victim] had cancelled a number of regular engagements scheduled for the week after his death." Responses were provided on a 7-point scale ranging from -3 (strong evidence that [the suspect] is guilty) to 0 (neither guilty nor innocent) to +3 (strong evidence that [the suspect] is innocent). It is important to note that evidence from Yearsley and Pothos (2016) was designed to be weak, but in this experiment the evidence was designed to range from strong to weak, indicating various degrees of the suspect's guilt and innocence.

4.2.1.1.2 **Procedure**

After providing informed consent, participants answered some basic demographic questions and were then provided with some instructions regarding a hypothetical murder scenario (originally employed in Tetlock, 1983, p.287, but we utilized the adapted version in Yearsley & Pothos, 2016, p.4):

"Mr. Smith has been charged with murder. The victim is Mr. Dixon. Smith and Dixon had shared an apartment for nine months up until the time of Dixon's death. Dixon was found dead in his bed, and there was a bottle of liquor and a half-filled glass on his bedside table. The autopsy revealed that Dixon died from an overdose of sleeping pills. The autopsy also revealed that Dixon had taken the pills sometime between midnight and 2 am. The prosecution claims that Smith slipped the pills into the glass Dixon was drinking from, while the defense claim that Dixon took deliberately took an overdose."

Participants were then asked a series of questions regarding the scenario, such as who the accused was or the time when Dixon ingested the pills etc., and they could not progress onto the rest of the survey until all questions were answered correctly. This was to reinforce the participant's memory of the scenario. Participants then rated the strength and verdict direction of the evidence. All forty statements were randomized and presented together. Participants were also informed that some pieces of evidence would be contradictory (as detailed above). As such, we told participants to rate the strength of each piece of evidence simply as they read them and independently of any other pieces of evidence. Once all statements were rated, participants were provided with a debrief.

4.2.2 Results

Table 4.2 presents the descriptive statistics for the evidence statements used in this pilot. Transformations were then conducted so that evidence strength could be viewed as a probability of a given verdict (0 being guilty; 1 being innocent; and .5 indicating neutral verdicts). Future experiments can then employ these probabilities to identify the statements which can induce the maximum violation of the TB inequality.

However, before exploring this, we first need to consider the questions we would need to use to elicit the correlations in the TB inequality, and whether these verdict measurements can be conducted in a way that they are not 'disturbing' (in the specific sense outlined above). Specifically, are differences in verdicts between two days, (difference between verdict at day 1 and verdict at day 1,2) equivalent to asking how a verdict changed across two time points (directly asking for the difference)? If the answer to this question is yes, then we can employ a single 'change' measurement in lieu of two separate measurements (it is the latter procedure which is fraught with possibility of being disturbing). If the answer is no, then the change question we employ somehow disturbs the relevant cognitive processing and renders the change measurement procedure a dubious way to compute TB correlations.

Table 4.2 Descriptive statistics and transformed values (from -3 to 3 to 0 to 1) for evidence statements used in the pilot study. Transformation of the means corresponds to 0 being guilty and 1 being innocent

	Evidence	Mean Strength	SD Strength	Transform
Statement 1	The bottle of sleeping pills had Smith's finger prints on it.	-1.69	1.04	0.2180
Statement 2	The local chemist testified that Smith had bought the sleeping pills in his pharmacy a month before Dixon died.	-1.47	1.02	0.2542
Statement 3	Smith was spotted on CCTV near the flat at around 3am. He seemed distressed and anxious.	-1.34	0.98	0.2761
Statement 4	Smith's fingerprints were found on the bottle of liquor at Dixon's bedside.	-1.25	1.17	0.2912
Statement 5	One of Smith's previous housemates reported that Smith made him feel threatened.	-1.21	1.11	0.2988
Statement 6	Smith had a previous conviction for violent disorder.	-1.07	1.31	0.3224
Statement 7	Neighbors reported overhearing Dixon and Smith engaged in heated conversations on several occasions during the previous month.	-0.93	0.81	0.3451
Statement 8	Dixon had no history of depression or related conditions.	-0.65	1.32	0.3923
Statement 9	Dixon had arranged a number of social engagements for the week after his death.	-0.65	0.95	0.3923
Statement 10	Friends and colleagues reported that Dixon did not seem obviously stressed or depressed in the days leading up to his death.	-0.62	1.04	0.3965
Statement 11	DNA from Smith was found at the crime scene.	-0.61	1.04	0.3981
Statement 12	A mobile phone tower near the flat pinged Smith's phone near the flat around 1am.	-0.51	1.20	0.4150
Statement 13	Dixon appeared to have a large quantity of savings.	-0.49	0.86	0.4175
Statement 14	Dixon was successful in his career and had recently been promoted.	-0.46	1.19	0.4226
Statement 15	The empty bottle of sleeping pills was found in the kitchen.	-0.41	1.00	0.4310
Statement 16	Dixon was engaged to be married.	-0.39	1.21	0.4343
Statement 17	The addition of the sleeping pills to the liquor was unlikely to have altered its taste.	-0.31	1.00	0.4478
Statement 18	An acquaintance of Smith and Dixon thought he saw Smith enter the building where his apartment is located at around 1am, however he could not make a positive identification.	-0.17	0.73	0.4722
Statement 19	An acquaintance of Smith and Dixon thought he saw Smith leave the building where his apartment is located shortly before midnight, however he could not make a positive identification.	-0.04	1.19	0.4941
Statement 20	There were no photographs of Smith and Dixon near the time of death.	0.05	1.07	0.51
Statement 21	A Doctor testified that at least three or four of the pills would have had to be consumed for an overdose.	0.18	0.94	0.53
Statement 22	Dixon had been single for many years.	0.34	1.12	0.56
Statement 23	Dixon had recently been made redundant.	0.42	1.25	0.57

Statement 24	A Doctor testified that taking only two or three of the pills would have been sufficient for an overdose.	0.45	1.06	0.58
Statement 25	Smith was not seen on any CCTV footage that evening.	0.53	1.07	0.59
Statement 26	The empty bottle of sleeping pills was found in Dixon's bedroom.	0.53	0.96	0.59
Statement 27	Dixon was several weeks in behind in rent payments.	0.59	0.95	0.60
Statement 28	The addition of the sleeping pills to the liquor would likely have slightly altered the taste.	0.61	0.88	0.60
Statement 29	Although Smith's DNA was found at the crime scene, this is to be expected since he was living at the flat at the time of Dixon's death.	0.62	1.15	0.60
Statement 30	There were a number of photographs of Dixon and Smith laughing and enjoying each other's company near the time of death.	0.63	1.23	0.61
Statement 31	A mobile phone tower, many miles away from the flat, pinged Smith's phone around midnight.	0.72	1.01	0.62
Statement 32	Smith had no previous criminal convictions.	0.81	1.15	0.63
Statement 33	Friends and colleagues reported that Dixon seemed anxious and distracted in the days leading up to his death.	0.95	1.03	0.66
Statement 34	One of Smith's previous housemates testified to his good character, saying he was generally kind and conscientious.	0.97	1.07	0.66
Statement 35	The local chemist testified that Dixon had bought the sleeping pills in his pharmacy a month before he died.	1.02	0.66	0.67
Statement 36	Neighbors reported that Dixon and Smith appeared to get on well and were unaware of any serious arguments or disagreements.	1.06	1.11	0.68
Statement 37	Smith's fingerprints were not found on the bottle of liquor.	1.23	1.38	0.70
Statement 38	Dixon had a history of depression, although not of attempted suicide.	1.37	1.46	0.73
Statement 39	Dixon had cancelled a number of regular engagements scheduled for the week after his death.	1.37	1.27	0.73
Statement 40	The bottle of sleeping pills had only Dixon's finger prints on it.	1.53	1.05	0.76

4.3 Control Experiment 1

This experiment consisted of two parts. The first part required participants to answer a verdict change question, asking whether the second day of evidence changed their answer relative to the first day of evidence. The second part consisted of asking participants to respond to individual (marginal) verdict questions after reading day 1 evidence and after reading the evidence of day 1, 2. Note, the aim of this control was to establish whether asking a change question equated to the same verdict change as the difference between verdicts at day 1 and 2. A within-subjects design was employed so the responses could be compared across the same participants.

4.3.1 Method

4.3.1.1 Participants

Participants were recruited using Prolific and we restricted sampling to UK nationals. They were paid £1.25 for their involvement. Sample size was set a priori to 50 participants (25 males and 25 females). Participants were between 19 and 88 years old (M_{Age} = 38.22 years old, SD = 14.76). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with most participants reporting 5 (*n* = 45) and several others (*n* = 5) reporting 4 or lower.

4.3.1.2 Materials and Procedure

After providing consent, participants answered some demographic questions and were then provided with some initial instructions regarding the task. Participants then read a scenario involving a hypothetical murder (note, this scenario is different from the Smith paradigm employed in the pilot):

"Mr. Snyder has been charged with murder. The victim is Mr. Jones. They were working together in the same company and one of their colleagues reported that Mr. Jones owed Mr. Snyder money. They both work in very stressful roles. Mr. Jones has been found outside his apartment with two stabbing wounds in the chest. The murder weapon, a pocketknife, was found in a bin a few meters down the road. The autopsy revealed that Mr. Jones died immediately as the wounds were found to have punctured his heart. It is estimated that time of death was between 8pm and midnight."

After reading, participants were reminded to keep an open mind and consider the defendant neither guilty nor innocent. They were then told they would undertake the role of a juror and read some evidence about the case across two trial days. Participants then read two pieces of prosecution evidence on day 1 of the trial, and then two pieces of defense evidence on day 2.

We had two variations of change questions tested in this control. Participants were therefore split into two conditions, whereby half would provide their verdicts immediately after reading the second day evidence and the others would complete a filler task before providing their verdict. We created variations of the change question procedure for exploratory purposes, as a way to compare immediate memory with slightly delayed memory. For the immediate change question, participants responded to: "Did the evidence on Day 2 change your mind, relative to before receiving the information?" Participants who answered the delayed change question responded to: “How much has your opinion changed between after receiving the Day 1 evidence and after receiving the Day 2 evidence?” Both questions were on a 0-100 slider scale (0 being a lot towards guilt; 100 being a lot towards innocent; and 50 being not at all).

Participants then completed a two-minute filler task before reading a near-identical scenario, with minor changes (e.g., names of the suspect and victim or the murder weapon, changed from a shard of bottle glass to a penknife, were altered slightly to create the impression of a similar, albeit different, case). This near-identical scenario corresponded to the verdict measurements asked after seeing the evidence for each day, described next.

After reading a scenario analogous to the one detailed above, participants read the evidence of day 1. They then answered the question: “What is your verdict of Mr. Miller, based on the evidence you have seen so far?”, on a slider scale from 0 (guilty) to 100 (innocent). Participants then read another analogous scenario and evidence across both days and then responded to the final verdict question, “What is your verdict of Mr. Hill, based on the evidence you have seen so far?”, via the same 0-100 scale. Once this second question was answered, participants were provided with a debrief.

4.3.2 Results

We refer to the verdicts provided after each individual day as ‘marginal scales’. Since these marginal scales corresponded to verdicts and the change scale to the degree of change in verdict, we needed to standardise the responses by converting both into change verdicts, to make them comparable. To do this, we first calculated the difference between the marginal verdicts at day 2 and day 1. This provided us with a direct measure of how participants’ verdicts changed from viewing the evidence at day 1 to viewing evidence after reading day 2. Next, we subtracted the no change score (50) from the change verdict given by the participant. Verdicts indicating a positive number indicated a change towards innocence (e.g. a verdict of 80 became 30 after subtracting 50, the no change score) whereas negative numbers indicated a change towards guilt (e.g., 40 became -10). Having done this, the difference between the two individual verdicts is equivalent to the change verdict.

Table 4.3 shows the raw marginal and change verdicts. Table 4.4 shows the adjusted verdicts by subtracting 50 from each of the marginal and change verdicts. The day 1 marginal indicates a minor inclination towards guilt (-4.58) whereas the day 2 marginal shifts towards a more innocent verdict. A simple subtraction of the day 1 measurement from the day 2 measurement reveals the innocent oriented marginal difference (10.62). We then conducted two paired samples t-tests to test the differences between the differences of marginal verdicts and each of the change verdicts. Neither the delayed change verdict nor the immediate change verdict was statistically significant from the marginal differences, $t(24) = -1.54$, $p = .137$ and $t(24) = -1.65$, $p = .112$, respectively.

Taken together, these results suggest similarity between our marginal and change verdicts, that is, the change measurements are equivalent to the difference between the two individual verdicts. However, note that we could not identify support for the null, using Bayesian paired samples t-tests ($BF_{01} = 1.68$; $BF_{01} = 1.45$, respectively).

Table 4.3 Raw verdicts for the marginal and change questions in the first control experiment.

	Day 1 marginal	Day 2 marginal	Change verdict: immediate	Change verdict: delayed
Mean	45.42	56.04	50.4	50.72
SD	18.65	20.61	21.36	18.41
N	50	50	25	25

Table 4.4 Adjusted verdicts for the marginal and change questions in the first control experiment. Note, Table 4.3 mean raw scores were each subtracted by 50.

	Day 1 marginal	Day 2 marginal	Marginal difference*	Change verdict: immediate	Change verdict: delayed
Mean	-4.58	6.04	10.62	.4	.72
N	50	50	50	25	25

*Computed by subtracting the mean verdict of day 1 from the mean verdict of day 1,2 (day 1,2 – day 1)

4.4 Control Experiment 2

This experiment was identical to the previous experiment, except for two amendments. The first amendment related to how we probed for change. The second amendment related to the use of the same change question across both immediate and delayed change conditions.

4.4.1 Method

4.4.1.1 Participants

Participants were recruited using Prolific and we restricted sampling to UK nationals. They were paid £1.25 for their time. As per the first control experiment, we recruited 50 participants (25 males and 25 females). Participants were between 19 and 71 years old (M_{Age} = 35.36 years old, SD = 13.63). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with nearly all participants reporting 5 (n = 48) and a couple of participants (n = 2) reporting 4 or lower.

4.4.1.2 Materials and Procedure

As noted, the procedure of this experiment was nearly identical to that of the first control experiment, but for two amendments. First, both conditions for the change question employed the same question. In the previous experiment, each condition involved a different question. Utilising the same change question for both conditions allows us to attribute any change results to the degree of immediacy corresponding to when participants offer their verdict. The second amendment relates to the introduction of a new change question. Specifically, participants were asked the question: "Think of how guilty you considered the suspect with the day 1 evidence. Has the day 2 evidence made you to consider him more or less guilty?" on the same 0 to 100 scale. Results from the first control experiment revealed little support, using Bayesian t-tests, for the acceptance of non-disturbing measurements across the marginal and change questions. One explanation for the results was that participants were unsure how to respond to the question, leading to increased noise. Specifically, participants answering the change question may have offered their verdict of the scenario, rather than provide a response on how their verdict specifically changed from the first day of evidence to the second. Thus, since neither change question from the first control experiment revealed equivalence with the difference between marginals, we employed a different change question to enhance the clarity of what we were asking participants to do.

4.4.2 Results

Once again, we first computed the difference between the marginal verdicts as well as the directional verdict change. We then conducted two paired samples t-tests to test the differences between the computed marginal differences and each of the change verdicts. Both the delayed change verdict and the immediate change verdict were significantly compared to the difference between marginals, $t(24) = -2.66, p = .014$ and $t(24) = -2.67, p = .013$, respectively. These results suggest no equivalence between the marginal and change question verdicts.

Next, we show the descriptive statistics for the main verdicts and verdict differences in this experiment (see Table 4.5). We first note the expected directional change between the marginals, whereby verdicts became more innocent after reading the evidence on day 2. However, a discrepancy emerges when we consider the change verdicts. According to both immediate and delayed conditions, participants' verdicts became more guilty, even though the evidence presented to participants was for innocence (see Table 4.6).

How do we explain these findings? Once again, we speculate that participants provided their verdicts of the suspect after reading the evidence on day 2 rather than providing the change in their verdict across both days. Our reasoning for this impression is based on the similarities between the individual marginals and the change verdicts (see Table 4.5). It is therefore important that our next experimental attempt simplifies the response process even further to prevent conflation between specific verdicts and verdict change.

Table 4.5 Verdicts for the marginal and change questions in the second control experiment.

	Day 1 marginal	Day 2 marginal	Change verdict: immediate	Change verdict: delayed
Mean	39.84	44.06	40.32	43.28
SD	18.92	18.3	24.01	20.15
N	50	50	25	25

*Computed by subtracting the mean verdict of day 1 from the mean verdict of Day 1,2 (Day 1,2 – Day 1)

NB. Compare the Day 2 marginal verdict with the change one: they are nearly identical, reinforcing our impression that participants mistook the 'change' question with a request to simply provide a verdict at Day 2.

Table 4.6 Adjusted verdicts for the marginal and change questions in the second control experiment. Note, Table 4.5 mean raw scores were each subtracted by 50.

	Day 1 marginal	Day 2 marginal	Marginal difference*	Change verdict: immediate	Change verdict: delayed

Mean	-10.16	-5.94	4.22	-9.68	-6.72
N	50	50	50	25	25
*Computed by subtracting the mean verdict of day 1 from the mean verdict of day 1,2 (day 1,2 – day 1).					

4.5 Control Experiment 3

In this experiment, three amendments were applied to the paradigm: 1) simplification of the change question scale, 2) simplification of the question wording, and 3) the experiment was conducted between-subjects.

4.5.1 Method

4.5.1.1 Participants

Participants were recruited using Prolific and we restricted sampling to UK nationals. They were paid £0.62 for their involvement (note, payment was halved because the length of study was halved). Sample size was set to 100 participants (50 males and 49 females and 1 participant who self-identified as ‘other’). Participants were between 18 and 58 years old (M_{Age} = 30.71 years old, SD = 9.52). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with most participants reporting 5 (*n* = 95) and a several others (*n* = 5) reporting 4 or lower.

4.5.1.2 Materials and Procedure

As noted, we conducted a near-identical replication of the previous experiment, but with three amendments. First, we simplified the scale of the change question by replacing the 0 to 100 range with a -3 to 3 scale. Our reasoning for this change was scale clarity; that is, it makes more intuitive sense for zero to represent no verdict change than a middle scale point of 50. Second, we further simplified the change question by asking participants, “Consider your verdict after reading the first day of evidence, and then consider your verdict after reading the second day of evidence. Was there a change in your verdict?”, i.e., directly asking for a verdict change between day 1 and 2. Finally, we ran this experiment between-subjects. We wanted to ensure that participants were not suffering from response biases by having to do two variants of the same scenario twice. So, we ran the marginal and change questions separately for different participants. Note, the two change questions were also run between-subjects.

4.5.2 Results

Two independent samples t-tests were conducted to compare verdicts between the difference of the ‘marginal’ questions and each of the change questions. We found a significant difference when comparing the marginals to the delayed change question, *t*(73) = -2.42, *p* = .018, but not when comparing to the immediate change question *t*(73) = -.082, *p* = 935.

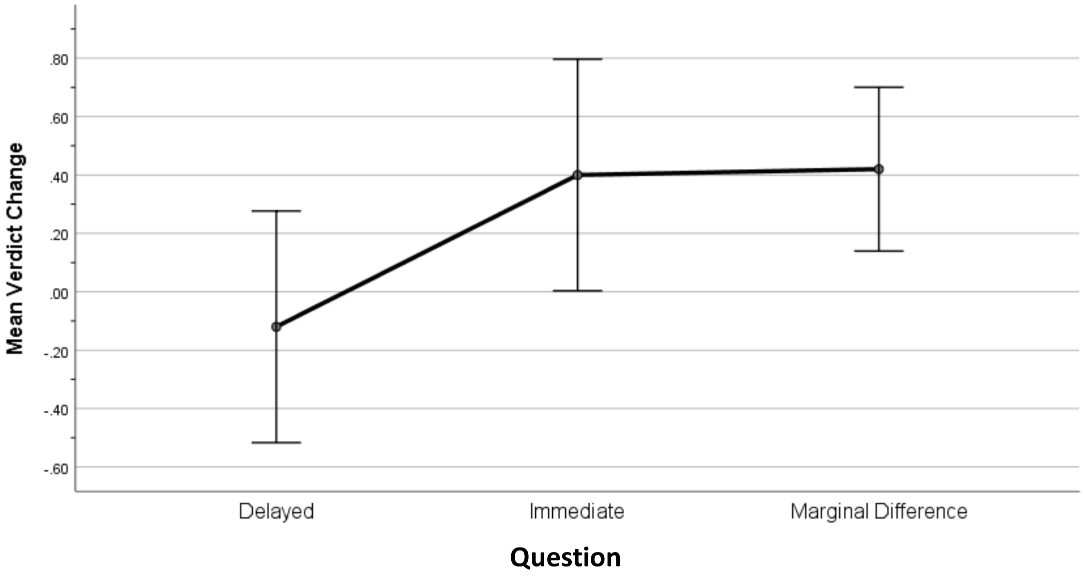
Next, we employed Bayesian independent samples t-tests to examine the degree of support for the null hypothesis. The delayed and marginal difference verdicts offered little support for the null hypotheses (BF$_{01}$ = .35), whereas the immediate change and marginal difference verdicts offered more convincing support (BF$_{01}$ = 3.98). These results, taken together with the impression from the descriptives in Table 4.7 and the data presented in Figure 4.4, suggest reasonable equivalence between the verdicts in the marginal questions and the immediate change question.

Table 4.7 Verdicts for the marginal and change questions in the third control experiment.

	Day 1 marginal	Day 1,2 marginal	Marginal difference*	Change verdict: immediate	Change verdict: delayed
Mean	-.58	-.16	.42	.4	-.12
SD	1.05	1.02	.86	1.22	1.01
N	50	50	50	25	25

*Computed by subtracting the mean verdict of day 1 from the mean verdict of Day 1,2 (Day 1,2 – Day 1).

Figure 4.4 Mean verdict change for the marginal and change questions in Control Experiment 3. Error bars refer to 95% confidence intervals.



4.6 Experiment 1

This experiment aimed to test whether the values of the questions across different pairs of time points (12, 23, and 13) constrain each other as required by the LGI. We employed the immediate change question from our third control experiment.

4.6.1 Method

4.6.1.1 Participants

Participants were recruited using Prolific and we restricted sampling to UK nationals. They were paid £1.16 for their contribution. Sample size was set to 150 participants (74 males and 75 females and 1 participant who self-identified as 'other'). Participants were between 18 and 70 years old (M_{Age} = 35.07 years old, SD = 12.84). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with the majority reporting 5 (n = 146).

4.6.1.2 Materials and Procedure

After providing informed consent, participants responded to some simple demographic questions and were then provided with some initial instructions. Participants then read the Smith scenario, as detailed in the pilot study, and were informed that the trial took place over three days. Each day corresponded to evidence indicating guilt, innocence, or neither. Note, mixed evidence was always presented on day 2, but whether guilty or innocent evidence was presented on day 1 or day 3 was counterbalanced. Two pieces of evidence were presented on each day, but the valence of the two pieces was selected (based on the pilot) to indicate either guilt or innocence or neither (this was based on a simple average of the valence of each piece of evidence on each day). Once all the evidence was presented, participants were asked only one of the three change questions: "Consider your verdict after reading the first day of evidence, and then consider your verdict after reading the third day of evidence. Was there a change in your verdict?" The other change questions corresponded to the change at time points 1,2 and 2,3. After answering the change question, participants were debriefed.

4.6.2 Results

In order to apply the LGI in a way analogous to how it is applied in physics (which involves binary observables), we first converted all scale change responses to their respective binary equivalents. In other words, a no change response was set to zero and a change response (responses from -3 to -1 and +1 to +3) were set to one. We next present Table 4.8, which shows the descriptive statistics for the verdict changes at each time pair. Since this change variable is on a scale of -1 to 1, we readily know that $C_{xy} = -\langle\Delta\rangle$. That is, recall, when we quantify $\Delta = +1$ for change and $\Delta = -1$, $C_{xy} = -\langle\Delta\rangle$; if change is encoded instead so that $\Delta' = +1$ for change and $\Delta' = 0$ for no change, we simply have $\Delta = 2\Delta' - 1$ and so $C_{xy} = -2\langle\Delta'\rangle + 1$.

At this point, it is possible to assess LGI violations. Even though there are different versions of the LGI (Halliwell, 2014) we restrict ourselves to a form canonical for the present problem, namely $C_{12}+C_{23}<C_{13}+1$. There was no evidence for a violation of the LGI (see Table 4.9).

Table 4.8. Descriptive statistics for change verdicts across each time pair in Experiment 1.

	Guilty first (*n* = 75)			Innocent first (*n* = 75)		
	Change 12	Change 23	Change 13	Change 12	Change 23	Change 13
Mean Change	.80	.92	.76	.60	.76	.84
SD	.41	.28	.17	.50	.44	.37
95% CI	.16	.11	.10	.20	.17	.15

Table 4.9 Transformed change verdicts across each time pair in Experiment 1. NB: Table 4.8 data was transformed using -2*change+1. Recall, if change is encoded instead so that $\Delta' = +1$ for change and $\Delta' = 0$ for no change, we simply have $\Delta = 2\Delta' - 1$ and so $C_{xy} = -2\langle\Delta'\rangle + 1$.

	Guilty first (*n* = 75)			Innocent first (*n* = 75)		
	Correlation 12	Correlation 23	Correlation 13	Correlation 12	Correlation 23	Correlation 13
-2*change+1	-.60	-.84	-.52	-.20	-.52	-.68

Table 4.10 Application of Table 4.9 data to the LGI in Experiment 1.

	Guilty first (*n* = 75)		**Innocent first (*n* = 75)**	
Leggett & Garg (1985)	**$C_{12} + C_{23} < C_{13} + 1$**			
	$C_{12} + C_{23} =$ -1.44	$C_{13} + 1 =$ .48	C12 + $C_{23} =$ -.72	$C_{13} + 1 =$ .32
TB broken?	No		No	

4.6.3 Discussion

Regardless of evidence order, we were unable to violate the TB inequality. One reason for this failure may have arisen from our use of scale questions. In this paradigm, we employed change questions so that no change responses rendered a zero response and change (either +1 to +3 or -1 to -3) were mapped to one. This ad-hoc adjustment is confusing as it meant that from the seven options made available (-3 to 3, including 0), six options indicated change and one option indicated no change. To address this issue, we present participants with binary responses (-1 = no change; +1 = change) in the next experiment. Additionally, this also adds some realism to the trial as participants would be thinking in terms of innocent versus guilty, just as jurors in the court room would be thinking.

4.7 Pilot 2

A major limitation of Experiment 1 was that we had to coarsen a variable from -3 to 3 into a binary one with {-1, 1} values. We address this problem next by directly employing binary questions. This would allow a more direct test of whether the LGI can be violated in this behavioral setup. The purpose of this pilot was to identify questions which would make it more likely that the LGI can be broken.

4.7.1 Method

4.7.1.1 Participants and Procedure

Three PhD students were presented with the Smith scenario and with a single piece of evidence across three days. The evidence statements used were based on the pilot experiment and were generally guilty in nature (see Tables 4.2 [above, pilot] and 4.11 [below]). Specifically, the evidence presented on days one and two were both individually weak guilty information, but, when combined with the stronger day 3 evidence, would be expected to create the right conditions to break one of the TB inequalities. Five different evidence combinations were created and tested (see Table 4.11 below). Each of the students was randomly allocated one of the time pairings for each of the evidence combinations. Participants were then asked, “Has your verdict (innocent versus guilty) changed between what it was on Day X and Day Y?” and would respond in a binary yes/no fashion. Participants were then asked to “briefly explain [their] choice”, whereby they were given the opportunity to freely express their reasoning for why their verdict changed or not. This was used for two reasons. First, it was to ensure that participants employed the appropriate time points in relation to the question they were meant to answer. For example, a participant using day 3 evidence when they should only be considering their verdict difference between days 1 and 2 would have misunderstood the instructions. Second, examination of the participants’ reasoning enables us to examine whether the intended story of the evidence was used when considering verdict change.

4.7.2 Results

Table 4.12 shows the verdicts provided by the PhD students in this study. Only one of the evidence combinations satisfied the requirements for breaking the LGI (i.e., no verdict changes across days 1,2 and 2,3, but a verdict change in days 1,3). As such, combination 1 will be employed in the second experiment.

Table 4.11. Evidence combinations as used in the second pilot experiment.

Evidence Combination #	**Day 1**	**Day 2**	**Day 3**

1	An acquaintance of Smith and Dixon thought he saw Smith enter the building where his apartment is located at around 1am, however he could not make a positive identification.	The empty bottle of sleeping pills was found in the kitchen.	The bottle of sleeping pills had Smith's fingerprints on it.
2	Dixon had arranged a number of social engagements for the week after his death.	Dixon had no history of depression or related conditions.	Smith's fingerprints were found on the bottle of liquor at Dixon's bedside.
3	Smith had a previous conviction for violent disorder.	One of Smith's previous housemates reported that Smith made him feel threatened.	Smith was spotted on CCTV near the flat at around 3am. He seemed distressed and anxious.
4	A Doctor testified that at least three or four of the pills would have had to be consumed for an overdose.	An acquaintance of Smith and Dixon thought he saw Smith leave the building where his apartment is located shortly before midnight, however he could not make a positive identification.	The addition of the sleeping pills to the liquor was unlikely to have altered its taste.
5	The empty bottle of sleeping pills was found in Dixon's bedroom.	The addition of the sleeping pills to the liquor was unlikely to have altered its taste.	DNA from Smith was found at the crime scene.

Table 4.12. Change verdicts across all time pairs in the second pilot experiment.

Evidence Combination #	**Change 1,2**	**Change 2,3**	**Change 1,3**	**TB violation?**
1	No	No	Yes	✓
2	No	Yes	No	✗
3	No	Yes	Yes	✗
4	No	No	No	✗
5	Yes	No	No	✗

4.8 Control Experiment 4

In this experiment, two amendments were made to the paradigm. First, binary questions were employed in both marginal and change questions. Second, RTs were measured to examine evidence that the change question did not just (psychologically) involve two separate judgements and a computational of their difference.

4.8.1 Method

4.8.1.1 Participants

Participants were recruited using Prolific and we restricted sampling to UK nationals. They were paid £1.25 for their involvement. Sample size was set to 100 participants (50 males and 50 females). Participants were between 19 and 66 years old (M_{Age} = 32.94 years old, SD = 11.51). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with all participants indicating extreme comfort with English.

4.8.1.2 Materials and Procedure

We conducted a near-identical replication of Control Experiment 3, but for the exception of two amendments. As noted, binary questions were employed in both marginal and change questions. In our previous control experiment, we established reasonable equivalence between scale change responses and marginal differences. Here, we attempt to find equivalence for when participants offer binary responses. The second amendment relates to the introduction of measurement for RTs. We included RTs to examine whether there was similar processing time across change and marginal questions.

4.8.2 Results

We first employed a preliminary data processing step, which consisted of identifying participants who spent too long responding to the marginal and change questions. Participants were removed if they spent longer than the mean RT plus 3 SDs. In total, we retained 97 participants and removed 3 participants (2 in the marginal condition and 1 in the change condition) from our analyses.

Next, we computed whether a change occurred between day 1 and 2 for each participant. This enabled us to compare the change verdicts for the marginal questions with the scores from the direct change question. For instance, if a participant provided an innocent verdict on both days, a zero would be coded for that participant (indicating no change in the verdicts between the days). However, if a participant rendered the suspect guilty on day 1 but innocent on day 2 (or vice versa), a one would be coded for that participant (indicating a change in the verdict between days).

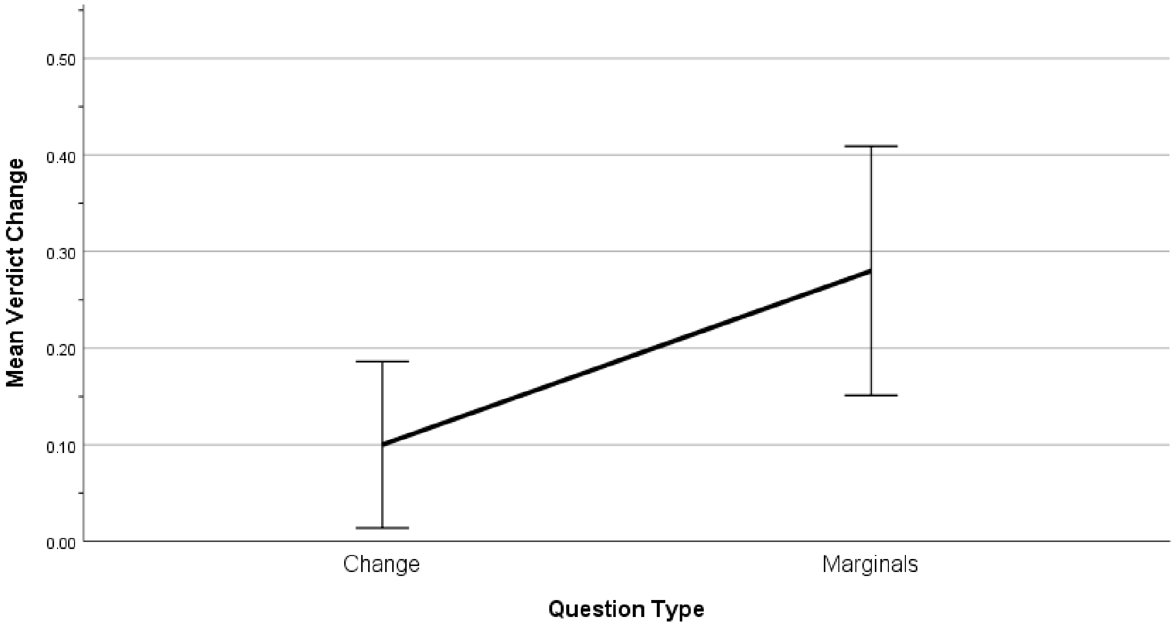
We then conducted an independent samples t-test to compare the change verdicts derived from the marginal questions and the change verdicts from the direct change question. We found a significant difference when comparing the marginals to the change question, $t(81.61) = -2.4$, $p = .019$ (note, Levene's test indicated unequal variances, $F = 26.67$, $p < .001$, and so degrees of freedom were adjusted from 95 to 81.61). Indeed, Table Z shows a notable difference when comparing the mean change from the difference of marginal questions (.29) to the mean direct change verdicts (.1). However, a Bayesian independent samples t-test indicated no support for the experimental hypothesis ($BF_{01} = .38$). We take these results, alongside the descriptives in Table 4.13 and the data displayed in Figure 4.5, to suggest little equivalence between the verdicts in the marginal questions and the immediate change question.

Table 4.13. Verdicts for the marginal and change questions in the fourth control experiment.

	Day 1 marginal	Day 1,2 marginal	Marginal change*	Change verdict
Mean	.65	.73	.29	.1
SD	.48	.45	.46	.31
N	48		-	49

* Derived from whether a change occurred for each participant. For instance, if a participant provided the same verdict on both days, a zero would be coded for that participant. If a participant's verdict differed across days, a one would be coded for that participant.

Figure 4.5. Mean verdict change for the change and marginal questions in Control Experiment 4. Error bars refer to 95% confidence intervals.



Finally, we consider RTs across the change and marginal questions. We first conducted a paired samples t-test between the RTs on marginal day 1 and marginal day 2. There was a non-significant difference ($t(47) = 1.39$, $p = .172$), suggesting participants took similar amounts of time to complete each of the marginals. We then conducted a series of independent samples t-tests to compare the RTs on marginal day 1 and change, marginal day 2 and change, and, lastly, change and the sum of RTs from both marginal questions. We conducted this latter comparison because we wanted to examine whether the processing time would be similar across all questions. Specifically, a participant considering their change question may require more time if they first need to consider their verdict at day 1, then at day 2, and then consider whether there was a difference between them.

Respectively, comparing each individual marginal RT to the change RT resulted in significant differences (marginal day 1: F (1, 95) = 35.84, $p < .001$, and marginal day 2: F (1, 95) = 44.59, $p < .001$), suggesting that participants took significantly longer to respond to the change question than they did for each of the marginals (see Table 4.14). Interestingly, this was not the case when comparing the sum of the RTs from each of the marginal questions to the change RTs. In fact, we find no significant difference in this case (F (1, 95) = .74, p = .39). We then employed a Bayesian independent samples t-test to examine the degree of support for the null hypothesis, which was found to be of weak-moderate support (BF_{01} = 3.37).

Table 4.14 RTs across each question for Control Experiment 4.

	Marginal Day 1	Marginal Day 12	Marginal Day 1 + Day 12	Change
Mean RT (in seconds)	6.3	5.24	11.54	12.55
SD	3.21	3.96	4.89	6.49
n	48			49

4.9 Experiment 2

This experiment was a direct replication of Experiment 1, but with various amendments (see procedure for more details). These amendments were applied to align closer the behavioural paradigm to the analogous physics set up.

4.9.1 Method

4.9.1.1 Participants

Participants were recruited using Prolific and we restricted sampling to UK nationals. They were paid £1 for their participation. Sample size was set to 100 participants, with 101 participants actually recruited (50 males and 51 females). Participants were between 18 and 58 years old (M_{Age} = 32 years old, SD = 10.07). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with the majority reporting 5 (*n* = 99) and the others (*n* = 2) reporting 4 or lower.

4.9.1.2 Materials and Procedure

The procedure was as in Experiment 1, except for the following amendments. First, in Experiment 1, participants responded to the change question on a -3 to +3 scale. However, in this experiment, participants answered the question with binary responses coded as: “Yes my verdict has changed” or “No my verdict remains the same”. Second, after providing their response, in a separate screen, participants were asked, “How did you determine whether there was a change verdict or not?” This question served as a manipulation check to ensure that participants were using the correct evidence, when considering their verdict changed. Finally, we employed the evidence combination from Pilot 2. Specifically, the evidence we employed was, "An acquaintance of Smith and Dixon thought he saw Smith enter the building where his apartment is located at around 1am, however he could not make a positive identification", "The empty bottle of sleeping pills was found in the kitchen", and "The bottle of sleeping pills had Smith's fingerprints on it", for days 1 to 3, respectively.

4.9.2 Results

Table 4.15 shows the descriptive statistics for the verdict changes at each time pair. Once again, we employ only the canonical LGI, $C_{12+}C_{23}<C_{13}+1$. There was no evidence for a violation of the LGI (see Table 4.17). In transforming the data from change scores to correlations, recall that, when we quantify $\Delta = +1$ for change and $\Delta = -1$, $C_{xy} = -\langle\Delta\rangle$; if change is encoded instead so that $\Delta' = +1$ for change and $\Delta' = 0$ for no change, we simply have $\Delta = 2\Delta' - 1$ and so $C_{xy} = -2\langle\Delta'\rangle + 1$. In this case, because change scores were quantified as +1 or 0, we employed the latter formula.

Table 4.15 Descriptive statistics for change verdicts across each time pair in Experiment 2.

	Change 12 (*n* = 34)	Change 23 (*n* = 33)	Change 13 (*n* = 34)
Mean Change	.12	.45	.5
SD	.33	.51	.51
95% CI	.11	.17	.17

Table 4.16 Transformed change verdicts across each time pair in Experiment 2. NB: Table 4.15 data was transformed using -2*change+1.

	Correlation 12	Correlation 23	Correlation 13
-2*change+1	.76	.09	0

Table 4.17 Application of Table 4.16 data to the LGI in Experiment 2.

Leggett & Garg (1985)	**$C_{12} + C_{23} < C_{13} + 1$**	
	$C_{12} + C_{23} =$ .86	$C_{13} + 1 =$ 1
TB broken?	No	

4.9.3 Discussion

This experiment made a handful of amendments to the paradigm, particularly in the adoption of a binary response format and employing the evidence indorsed by Pilot 2. Nevertheless, the LGI was still not violated. One reason why we failed to violate the LGI might have been because participants were confusing their change verdicts across the different time points. Specifically, there are only two relevant change periods, but we have three pieces of evidence. So, one of those pieces of evidence does not contribute to a 'change', it only sets the initial state.

Let us consider the current change rotations, that is, C12: change from *after* hearing evidence on day one, to *after* hearing evidence on day two (i.e., change caused by day two evidence alone, not day one), C23: change from *after* hearing evidence on day two, to *after* hearing evidence on day three (i.e., change caused by day three evidence alone, not day two); and, C13: change from *after* hearing evidence on day one, to *after* hearing evidence on day three (again, not any change caused by evidence on day one). In other words, the evidence on day 1 does nothing because all decisions are made relative to belief after hearing it.

Instead, we propose recasting the paradigm with time 1 as the initial state (innocent verdict prior to any evidence), with a piece of evidence for times 2 and 3. Note, times 2 and 3 are hereafter referred to as evidence days 1 and 2, respectively.

4.10 Pilot 3

The paradigm so far has employed three pieces of evidence across three trial days. This may have been problematic if participants had trouble keeping track of which piece of evidence was presented on different days. We address this problem by recasting the paradigm with two trial days and an initial verdict of innocence (to conform to the principle of 'innocent until proven guilty) prior to any evidence. The present pilot was used to identify evidence which be suitable for a test of the LGI.

4.10.1 Method

4.10.1.1 Participants and Procedure

Four PhD students were presented with the Smith scenario and were asked to initially consider Mr. Smith as innocent, prior to the presentation of any evidence. This pre-trial verdict is analogous to the day 1 timepoint from previous experiments but differs in the sense that we aim for all participants to set their initial state to innocent. The students were then given a single piece of evidence across two days (time points 2 and 3). Note, the evidence used in this experiment was derived from the initial pilot experiment and generally indicated guilt (see Tables 4.2 [above, pilot] and 4.18 [below]). Specifically, the evidence presented on day one (time point 2) comprised of weak guilty information, followed by strong guilty evidence on day 2. Recall, our reasoning for this was to create the right conditions to break the LGI inequality. As such, four different evidence combinations were created and tested (see Table 4.18 below). Each of the participants was then randomly allocated to one of the time pairings for each of the evidence combinations.

After reading the evidence, participants were asked a change question, e.g., “Did you change your verdict between the start and after day 2 evidence?” and responded in a binary manner. Note, there were three variations of the change question. In addition to the above change question, participants were also asked whether they had changed their verdict between the “start and after day 1 evidence”, as well as “after day 1 and after day 2 evidence”. Also, once again, we wanted to ensure that the participants’ reasoning matched the intended story of the evidence, and so they were asked, “How did you determine whether there was a change in your verdict or not?”.

4.10.2 Results

Table 4.19 shows the verdicts provided by the participants in this study. Only one of the evidence combinations satisfied the requirements for breaking the LGI (i.e., no verdict changes across time points 1,2 and 2,3, but a stronger change across time point 1,3). As such, combination 2 will be employed in the third experiment.

Table 4.18 Evidence combinations as used in the third pilot experiment. Note, the initial state (time point 1) was set to innocent, and so only two days of evidence were presented.

Evidence Combination #	**Day 1**	**Day 2**
1	The empty bottle of sleeping pills was found in the kitchen.	The bottle of sleeping pills had Smith's fingerprints on it.

2	The addition of the sleeping pills to the liquor was unlikely to have altered its taste.	Smith's fingerprints were found on the bottle of liquor at Dixon's bedside.
3	Dixon had no history of depression or related conditions.	The local chemist testified that Smith had bought the sleeping pills in his pharmacy a month before Dixon died.
4	DNA from Smith was found at the crime scene.	Smith was spotted on CCTV near the flat at around 3am. He seemed distressed and anxious.

Table 4.19 Change verdicts across all time pairs in Pilot 3.

Evidence Combination #	**Change 1,2**	**Change 2,3**	**Change 1,3**	**TB violation?**
1	No	No	No	×
2	No	No	Yes	✓
3	Yes	No	Yes	×
4	Yes	No	Yes	×

4.11 Experiment 3

This experiment offers a further test of the LGI based on Pilot 3. RTs were also recorded to compare how long participants would take to respond on the change question versus the questions for each individual day. Detailed explanations of what a 'change' measurement is were also added to the experiment.

4.11.1 Method

4.11.1.1 Participants

Participants were recruited using Prolific and we restricted sampling to UK nationals. They were paid £1.80 for their participation. For reasons outlined shortly, we collected our sample in two batches of 200 participants (due to online recruitment, we obtained an additional participant). In total, we recruited 401 participants (196 males, 201 females, 2 non-binary and 2 preferred not to say). Participants were between 18 and 73 years old (M_{Age} = 34.39 years old, SD = 11.39). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with most reporting 5 (*n* = 387) and the others (*n* = 14) reporting 4 or lower.

4.11.1.2 Materials and Procedure

The procedure was near identical to Pilot 3, except that we only employed evidence combination 2. Specifically, the evidence for day 1 was, "The addition of the sleeping pills to the liquor was unlikely to have altered its taste" and the evidence for day 2 was, "Smith's fingerprints were found on the bottle of liquor at Dixon's bedside." Note, in this experiment time point 1 was set to the initial innocent state with days 1 and 2 as time points 2 and 3, respectively.

Additionally, before participants read the Smith scenario, we provided some further instructions about the study. Specifically, we explained to participants what a 'change' judgement was, as well as showing a timeline picture which detailed what each change question was asking (for specific instructions, see Figure 4.6).

Figure 4.6 Change verdict instructions presented to participants in Experiment 3. Note, horizontal lines indicate page breaks in the Qualtrics survey.

You will shortly see some information about a suspect, Mr. Smith, accused of murdering a certain Mr. Dixon (of course, this is all hypothetical).

Mr. Smith is either guilty or innocent. However, Mr. Smith should initially be considered innocent.

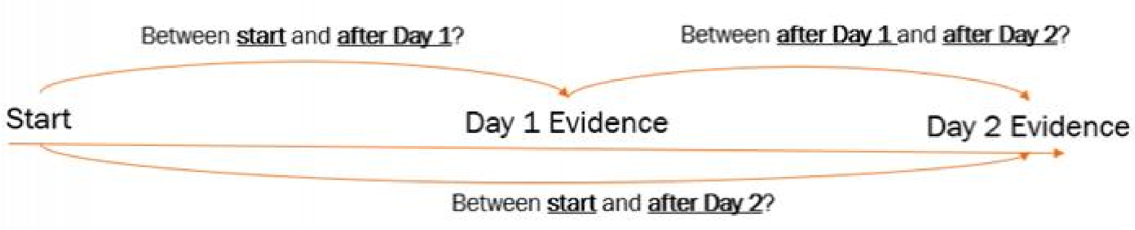
In this study we will ask you to make ‘change' judgments. **What is a 'change' judgment?**

Imagine being asked whether you are feeling colder or warmer. Sometimes you can answer such a question, without considering whether you are cold or warm in some absolute sense. That is, you could answer: I feel colder relative to 10 minutes ago.

You will be given evidence over two days of the trial. Your task is to determine whether there was a change in your verdict.

Now consider the timeline picture below.

In this experiment, you will be presented with some evidence and then you will be asked: "Would you say you changed your mind about Smith’s guilt..." (and one of the three questions in the picture below).


Between start and after Day 1?
Between after Day 1 and after Day 2?
Start
Day 1 Evidence
Day 2 Evidence
Between start and after Day 2?

'After Day 1' means after you have seen the Day 1 evidence.

'After Day 2' means after you have seen the Day 2 evidence.

So, after Day 2, you would have seen both the Day 1 and the Day 2 evidence (i.e., there is no more evidence).

Please take a moment to ensure you understand what will be asked of you.

4.11.2 Results

4.11.2.1 Data Screening

A preliminary data processing step was conducted, which corresponded to a minimal test of whether participants correctly remembered which piece of evidence corresponded to each day. Recall, this is a crucial requirement for the experimental design to offer a valid test of the LGI. Immediately after providing their change verdicts, participants were asked two questions: 1) “What was the evidence presented to you on Day 1?” and 2) “What was the evidence presented to you on Day 2?”. Participants were excluded if they failed to provide the appropriate evidence on the correct day. The answers were evaluated on the basis of a majority rule, consisting of three independent assessors, all broadly familiar with this work. In total, we retained 289 participants and excluded 112 participants from our data analyses.

4.11.2.2 Data Analyses

Table 4.20 shows the descriptive statistics for the verdict changes at each time pair. Once again, we employ only the canonical LGI, $C_{12+}C_{23}<C_{13}+1$. The statistical test in this case involved comparing the change counts across the 12 and 23 periods against the change counts across the 13 period; recall the relevant form of the TB inequality is $N_{-}(t_1, t_3) < N_{-}(t_1, t_2) + N_{-}(t_2, t_3)$. Once again, there was no indication that the LGI was violated (see Table 4.23). There was no statistical difference between verdict change in C12,C23 and C13, χ^2 = 1.814*, p* = .178. Note, the counts of C12 and C23 were combined, because this is what is required for the test of this form of the TB inequality.

Table 4.20. Frequency table for change question and verdict change in Experiment 3.

		Verdict		Total
		No Change	Change	
C12 & C23	OC	183	12	195
	EC	180.2	14.8	
C13	OC	84	10	94
	EC	86.8	7.2	
Total	OC	267	22	289
	EC			

*OC = Observed count; EC = Expected count

We now turn to the RTs across the change and marginal questions. We first conducted a one-way repeated measures ANOVA on the change and marginal (day 1 and day 12) RTs, which was significant, *F* (2, 576) = 6.886, *p* = .001, η^2 = 023. We then conducted a series of paired samples t-tests to compare change and marginal day 1, change and marginal day 12, and finally marginal day 1 and marginal day 12. There were no significant differences between the change and marginal day 1 RTs, *t*(288) = .813, *p* = .417. However, we did find statistically significant differences between the change and marginal day 12 RTs, as well as between each of the marginal trial RTs, *t*(288) = 3.506, *p* = .001, *t*(288) = 2.885, *p* = .004, respectively; see Table 4.24 for descriptive statistics). However, note that we identified only moderate support for the change and marginal x null, using Bayesian paired samples t-tests

(BF_{01} = 10.94), but no support for the change and marginal x,y, and marginal x and xy nulls (BF_{01} = .04; BF_{01} = .26, respectively; note, this last comparison is of less interest).

Table 4.21 Descriptive statistics for change verdicts across each time pair in Experiment 3.

	Change 12 (*n* = 105)	Change 23 (*n* = 90)	Change 13 (*n* = 94)
Mean Change	.02	.11	.11
SD	.14	.32	.31
95% CI	.03	.07	.06

Table 4.22 Transformed change verdicts across each time pair in Experiment 3. NB: Table 4.21 data was transformed using -2*change+1.

	Change 12	Change 23	Change 13
-2*change+1	.96	.78	.79

Table 4.23 Application of Table 4.22 data to the LGI in Experiment 3.

Leggett & Garg (1985)	**$C_{12} + C_{23} < C_{13} + 1$**	
	$C_{12} + C_{23}$ = 1.74	C_{13} + 1 = 1.79
TB broken?	No	

Table 4.24 RTs across each question for Experiment 3. Note, the change questions were between-subjects and the marginal questions were repeated measures.

	Change 12	Change 23	Change 13	Overall Change	Marginal Day 1	Marginal Day 12
Mean RT (in seconds)	5.88	8.60	8.65	7.63	7.23	5.96
SD	3.05	5.20	9.03	6.31	6.89	6.01

4.11.3 Discussion

Despite our latest amendments to the experimental paradigm, we were unable to break the LGI. It is worth highlighting how close we were to violating the LGI in this iteration. In previous experiments, we found the C12 and C23 terms fail to exceed the C13 terms by differences of 1.92, 1.04 (Experiment 1: guilty first, innocent first, respectively) and .14 (Experiment 2). In this experiment, we failed to violate the LGI by .05. Clearly, we are getting closer to the violation of the LGI, but what else can be done? Let us begin with the evidence. In Experiment 3, there was a near-identical rate of change for both t(2,3) and t(1,3) correlations. Recall, in order to violate the LGI, we require low rates of change in the t(1,2) and t(2,3) questions and a high rate of change in the t(1,3) question.

In the next experiment, we propose changing the day 1 evidence, t(2). Our reasoning for this is to increase the guilt rating of t(2) so to reduce verdict change between the two days of evidence (i.e., t(2,3)). As found in Table 4.2, the ratings of the evidence statements in Experiment 3 (day 1 evidence = -.31; day 2 evidence = -1.69; difference = 1.38), and the rating difference between the two evidence days were reduced in experiment 4 (day 1 evidence = -1.07; day 2 evidence = -1.69; difference = .62).

Next, we turn to the instructions incorporated into the paradigm. Experiment 3 presented detailed instructions regarding the change question, including the addition of a diagram to illustrate what is meant by each change question. In the next experiment, we further refine the instructions presented to participants. It is also worth highlighting that the diagram (shown in Figure 4.6) was only presented to participants in the initial phase of the experiment when the instructions were explained. In the next experiment, not only do we improve the aesthetic of the illustrations, but we also personalise the diagrams depending on the change question at hand; that is, if you were asked whether your verdict had changed between the initial verdict and day 1, the initial illustration would be presented but with highlighted boundaries on the initial verdict and day 1 time points.

4.12 Experiment 4

In this experiment, we refined the experimental setup of Experiment 3. Firstly, this experiment used a different crowdsourcing platform (Amazon's Mechanical Turk). Second, we improved the change instructions provided to participants. We also provided highlighted illustrations to ensure that participants understood what question and information was necessary in order to answer the question. Finally, we recast the evidence presented on day 1.

4.12.1 Method

4.12.1.1 Participants

Participants were recruited using Amazon's Mechnical Turk and we restricted sampling to US nationals who were deemed as 'Master workers'. Note, this term is awarded to participants that have demonstrated a high degree of success in performing a wide range of tasks on the platform. They were paid $2.39 for their participation (at the time of writing this was the exact conversion from our typical payment of £1.80). We recruited 201 participants (105 males, 95 females and 1 non-binary). Participants were between 24 and 71 years old (M_{Age} = 43.34 years old, SD = 10.45). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with every participant reporting extreme comfort in communicating in English (n = 201).

4.12.1.2 Materials and Procedure

We retained much of the procedure from Experiment 3, such as the initial verdict set to innocent followed by two evidence days, the same Smith scenario was presented, and identical wording of the three change questions, e.g., "Did you change your verdict between your initial innocent verdict and after reading the day 1 evidence?". However, this experiment refines Experiment 3 and includes some minor adjustments to violate the LGI.

Firstly, we changed the day 1 evidence. The day 1 evidence was changed from, "The addition of the sleeping pills to the liquor was unlikely to have altered its taste" to be "Smith had a previous conviction for violent disorder". Our reasoning for this was to increase the guilt rating of the day 1 evidence so to reduce verdict change between the two days of evidence (i.e., t(2,3)). As found in Table 4.2, the pilot ratings of the statements in Experiment 3 (day 1 evidence = -.31; day 2 evidence = -1.69; difference = 1.38), and the rating difference between the two evidence days were reduced in experiment 4 (day 1 evidence = -1.07; day 2 evidence = -1.69; difference = .62).

We also made some adjustments to the instructions presented before the Smith scenario. Specifically, we refined our explanation of what a 'change' judgement was, as well as showing an improved timeline picture that helped to explain what was required of the participant (for specific instructions, see Figure 4.7).

Figure 4.7. Change verdict instructions presented to participants in Experiment 4. Note, horizontal lines indicate page breaks in the Qualtrics survey.

You will shortly see some information about a suspect, Mr. Smith, accused of murdering a certain Mr. Dixon (of course, this is all hypothetical).

Mr. Smith is either guilty or innocent. However, Mr. Smith should initially be considered innocent.

In this study we will ask you to make ‘change' judgments. What is a 'change' judgment?

Imagine being asked whether you are feeling colder or warmer.

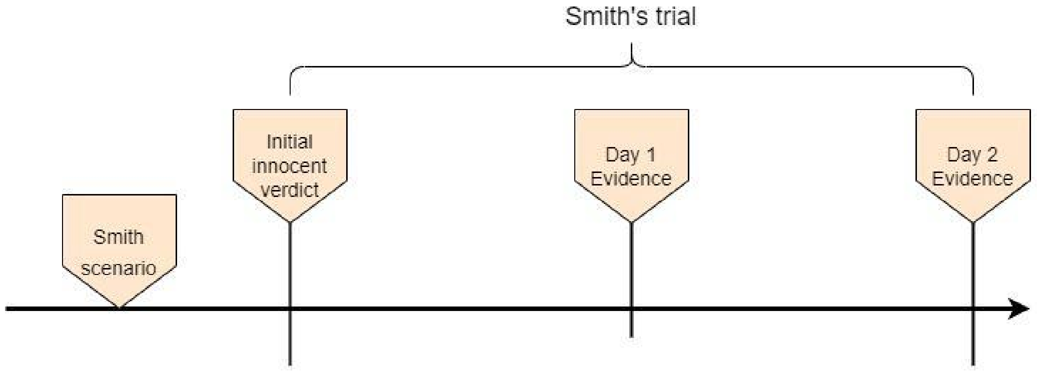
Sometimes you can answer such a question, without considering whether you are cold or warm in some absolute sense. That is, you could answer: I feel colder relative to 10 minutes ago.

So, for a change judgment, what matters is not the value, but the change: in this example, it does not matter whether you are feeling cold or hot, but rather if you are feeling colder or warmer, relative to 10 mins ago.

In the main part of the experiment, we will ask you at some point for a change judgment, for the guilt versus innocence of a hypothetical suspect. What matters is not whether you consider the person guilty or innocent, but whether your verdict has changed (for example, whether you considered the person initially innocent but now you think he is guilty; or the other way round!).

Now consider the timeline picture below.

You will first read a scenario regarding Mr Smith. This will provide you with context regarding Smith's trial.



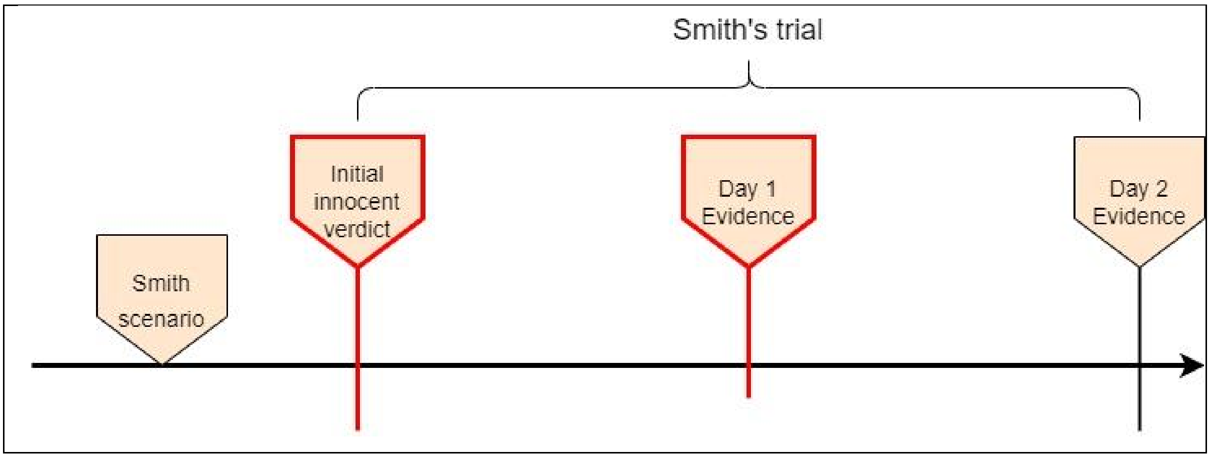
In criminal court, it is standard procedure to assume innocence of the suspect until they are proven guilty. At the beginning of this mock trial, we would like you to assume Mr Smith is innocent.

You will first read the Day 1 evidence, and then shortly after you will read the Day 2 evidence.

After reading the evidence you will be asked a 'change' question, where you will need to determine whether your verdict has changed between a specific period of time.

Figure 4.8 Example of a change question and illustration to ensure participant understanding in Experiment 4.

Did you change your verdict **between your initial innocent verdict** and **after reading the day 1 evidence**?



4.12.2 Results

4.12.2.1 Data Screening

A preliminary data processing step was conducted, which corresponded to a minimal test of whether participants correctly remembered which piece of evidence corresponded to each day. Recall, this is a crucial requirement for the experimental design to offer a valid test of the LGI. Immediately after providing their change verdicts, participants were asked two questions: 1) “What was the evidence presented to you on Day 1?” and 2) “What was the evidence presented to you on Day 2?”. Participants were excluded if they failed to provide the appropriate evidence on the correct day. The answers were evaluated on the basis of a majority rule, consisting of three independent assessors, all broadly familiar with this work. In total, we retained 172 participants and excluded 29 participants from our data analyses.

4.12.2.2 Data Analyses

Table 4.26 shows the descriptive statistics for the verdict changes at each time pair and Table 4.27 reveals the transformed change verdicts. For the first time in this experimental series, we provide evidence that the TB inequality can be behaviorally violated (see Table 4.28). Note, we also find a significant difference between verdict changes in C12, C23 and C13 (*p* = .017, two-tailed Fisher’s exact test*)*. Since the expected counts of 25% of cells were below five, we once again violated an assumption of the chi square test (see Table 4.25 below), and so Fisher’s exact test was employed.

Table 4.25 Crosstabs table for change question and verdict change in Experiment 4.

		Verdict		Total
		No Change	Change	
C12 & C23	OC	113	6	119
	EC	108.6	10.4	
C13	OC	44	9	53
	EC	48.4	4.6	
Total	OC	157	15	172
	EC			

*OC = Observed count; EC = Expected count

Once again, we examined RTs across the change and marginal questions (see Table 4.29). We first conducted a one-way repeated measures ANOVA on the change and marginal (x and xy) RTs, which was significant, F (2, 342) = 20.353, p < .001, η^2 = .11. We then conducted a series of paired samples t-tests to compare change and marginal x, change and marginal xy, and finally marginal x and marginal xy RTs. There was a significant difference between the change and marginal x, t(171) = 4.883, p < .001) and change and marginal xy, t(171) = 5.801, p < .001, but not between either of the marginals, t(171) = 1.824, p = .07; see Table 4.26 for descriptive statistics). We found little support for the null hypothesis for the comparison between marginal x and marginal xy, using Bayesian paired samples t-tests (BF_{01} = 2.325), and no support for the change and marginal x, nor the change and xy null (BF_{01} = 1.979^{-4}; BF_{01} = 3.162^{-6}, respectively).

Table 4.26 Descriptive statistics for change verdicts across each time pair in Experiment 4.

	Change 12 (*n = 64*)	Change 23 (*n = 55*)	Change 13 (*n = 53*)
Mean Change	.05	.05	.17
SD	.21	.23	.38
95% CI	.05	.06	.1

Table 4.27 Transformed change verdicts across each time pair in Experiment 4. NB: Table 4.26 data was transformed using -2*change+1.

	Change 12	Change 23	Change 13
-2*change+1	.91	.89	.66

Table 4.28 Application of Table 4.27 data to the LGI in Experiment 4.

	$\mathbf{C_{12} + C_{23} < C_{13} + 1}$

Leggett & Garg (1985)	$C_{12} + C_{23} =$ 1.8	$C_{13} + 1 =$ 1.66
TB broken?	Yes	

Table 4.29 RTs across each question for Experiment 4. Note, the change questions were between-subjects and the marginal questions were repeated measures.

	Change 12	Change 23	Change 13	Overall Change	Marginal Day 1	Marginal Day 12
Mean RT (in seconds)	7.24	9.11	9.36	8.57	5.99	4.91
SD	6.52	8.02	6.9	7.15	4.44	7.65

4.12.3 Discussion

The TB inequality was violated for the first time in this experimental series. This can be taken to be evidence against macrorealism, that is, the assumption that a question can have a specific value at all times. Without macrorealism, resolving a question is expected to change the relevant mental representations, so that order or interference effects would arise. The problem is that there are many kinds of ways in which a recollective process or judgement can impact on the relevant mental representations. A measurement that is too coarse (a 'sledgehammer' measurement) is likely to change the relevant system, but a corresponding conclusion would simply tell us that greater care is needed with our measurement approach. In physics, the NIM assumption is the one which tests whether a measurement is sufficiently adroit to prevent disturbance of a target system (Wilde & Mizel, 2012) or not. In psychology, we think we can avoid problems with the NIM assumption, by employing these 'change' judgements. However, there is a danger that a change judgement cognitively involves two individual judgements and their difference.

We utilised RTs to examine whether the change question does not just (psychologically) involve two separate judgements and a computation of their difference. However, in this experiment, processing times were dissimilar across the marginal and change questions. Indeed, RTs for both marginal questions (*M*=5.99, *M*=4.91) were significantly shorter than the RTs for the change question (average *M*=8.57). So, how do we reconcile the violation of TB with these results from RTs?

It is difficult to comment on the quantum-like nature of the TB violation without non-disturbing measurements. Let us instead focus on how the results concerning RTs may have arisen and how we can address this in a follow up experiment. Recall that in our experimental paradigm participants responded to the Smith scenario for the change questions and the Hill/Snyder scenarios for each of the marginal questions. It is possible that the observed differences in RTs could simply reflect practice effects. All participants responded to the scenarios in the same order, starting with the change question, then marginal 1, and then marginal 12. We would expect the most processing time to be taken by the earlier trials (due to novelty and understanding the response format) and the least processing time to be taken by the latter trials (when the participant is aware of what to expect and how they need to answer). Therefore, a simple solution is to employ the Smith scenario with the marginal questions, which would allow a measure of RTs for each marginal question, unadulterated by practice effects.

4.13 Experiment 5

This experiment was designed to test whether a change judgement involved two separate individual judgements or a single (well, change) judgement . This is important because a violation of the TB inequality can either occur because there is a quantum-like structure (violating macrorealism) and/or because in the two measurements that are needed for each term in the LGI (e.g, C12, C23, C13) the first measurement disturbs the system. We can examine whether a change judgement corresponds to two individual judgements or not by evaluating the differences between the RTs of the change question relative to the individual judgements.

4.13.1 Method

4.13.1.1 Participants

Participant recruitment was identical to that of Experiment 4, whereby Amazon's Mechanical Turk was used to recruit US national 'Master workers'. They were paid $1 for their participation. We recruited 100 participants (55 males, 44 females; 1 preferred not to say). Participants were between 28 and 78 years old (M_{Age} = 45.72 years old, SD = 10.2). Participants also reported their English fluency on a scale from 1 (extremely uncomfortable) to 5 (extremely comfortable), with almost every participant reporting extreme comfort in communicating in English (*n* = 99).

4.13.1.2 Materials and Procedure

We retained much of the materials and procedure from Experiment 4, including the Smith scenario and the evidence presented on days 1 and 2. Recall, that day 1 evidence was "Smith had a previous conviction for violent disorder" and day 2 evidence was "Smith's fingerprints were found on the bottle of liquor at Dixon's bedside". In previous experimental iterations, the Smith scenario was employed for the change question, and we used two different analogous scenarios involving the hypothetical suspects Mr. Hill and Mr. Snyder for the marginal verdicts for after day 1 and after days 1 and 2 (see Control Experiment 1). In this experiment, we employed the Smith scenario in the marginal questions. Specifically, we asked participants to provide a verdict after reading either the day 1 evidence or after reading both day 1 and 2 evidence. Importantly, we recorded RTs for each of the marginal questions. This meant that we could directly compare RTs from the change question in Experiment 4 to the marginal RTs in the current experiment.

4.13.2 Results

4.13.2.1 Data Analyses

We first conducted a one-way repeated measures ANOVA on the change and marginal (x and xy) RTs, which was non-significant, *F*(2, 271) = .334, *p* = .716 (see Table 4.30 for descriptive statistics). We then conducted a series of independent samples t-tests to compare change and marginal x, change and marginal xy, and finally marginal x and marginal xy. There were non-significant differences for all of the tests, *t*(220) = .433, *p* = .666; *t*(193.362) = 1.095, *p* = .275; *t*(81.196) = -.356, *p* = .723, respectively. Note, for the latter two tests, degrees of freedom had to be adjusted, because of significant Levene's tests.

Next, we conducted Bayesian independent samples t-tests for all the above comparisons. We found moderate evidence for all comparisons: marginal x and change RTs (BF_{01} = 5.3), marginal xy and change RTs (BF_{01} = 4.53), and the two marginals (BF_{01} = 4.48).

Table 4.30 RTs across each question for Experiment 5. Note, the change RTs were taken from Experiment 4 and the marginal RTs were from Experiment 5.

	Change 12	Change 23	Change 13	Overall Change	Marginal Day 1	Marginal Day 12
Mean RT (in seconds)	7.24	9.11	9.36	8.57	8.96	9.25
SD	6.52	8.02	6.9	7.15	4.91	3.01

4.13.3 Discussion

When examining the RTs between a change judgement and either of the two possible individual judgements, there is equivalence between the former and the latter. This result supports the assumption that a change judgement is *not* two individual judgements and a computation of a difference, which is the assumption needed to allow a conclusion of quantum-like structure from the violation of the TB inequality in Experiment 5.

4.14 General Discussion

The observed violation of the TB inequality can be taken to be evidence against macrorealism, that is, the assumption that a question can have a specific (even if unrevealed) value at all times. Without macrorealism, resolving a question is expected to change the relevant mental representations, so that order or interference effects would arise. In both the case of memory (Howe, 2011; Schacter et al., 2011) and more generally (e.g., Hogarth & Einhorn, 1992; Schwarz, 2007; Sharot et al., 2010), there is plenty of evidence of constructive processes, so what is the added value of an examination involving the unfamiliar framework of the TB inequality?

The problem is that there are many possible ways in which a recollective process or judgement can impact on the relevant mental representations. A measurement that is too coarse (a ‘sledgehammer’ measurement) is likely to change the relevant system, but a corresponding conclusion would simply tell us that greater care is needed with our measurement approach. In physics, the NIM assumption is one which tests whether a measurement is sufficiently adroit to prevent disturbance of a target system (Wilde & Mizel, 2012) or not. However, in physics it has been hotly debated whether it is at all possible to test for violations of the TB inequality, while conforming with the NIM assumption (Emary, 2017; Emary et al., 2015; Halliwell, 2016). In behavioral sciences, we believe we can get round this problem, by requesting change judgements, which can be directly related to the quantities needed to test for violations of a TB inequality.

So, with change judgements, the NIM assumption should be applicable, and therefore a violation of the TB inequality can be interpreted as evidence for quantum-like structure in human memory processes. For the first time, we offered empirical evidence showing such a violation, together with other unsuccessful attempts, illustrating that fine tuning is needed to obtain a TB inequality violation. Whether constructive processes in recollection are quantum-like versus not is a key issue, since the former possibility allows us to constrain the nature of such processes, in terms of the specific operations allowed in QT (cf. White et al., 2020). Quantum-like memory models and ideas have already been explored (Brainerd et al., 2013; Manning, 2021; Trueblood & Hemmer, 2017), though the use of the TB inequality offers a more general/ generic test of these ideas.

There is a more subtle way why evidence for quantum-like structure in memory is potentially significant. We can consider a recollection as a correlation between queries at present and memories for events in the past. We can then ask about the maximum correlation for particular mappings, between queries and events, if we assume classical resources (CPT) or quantum-like resources. As it turns out, in general, utilizing quantum-like resources leads to higher correlations, than with classical resources (Budroni et al., 2019). That is, a memory system with quantum-like representations may allow more efficient recollective processes. Although more work is needed to substantiate this proposal, the essential idea can be explained simply. Classically a conjunction can never be higher than a marginal, but behaviorally people often conclude that $p(A\&B) > p(B)$, a famous finding in both decision-making (Tversky & Kahneman, 1983) and memory (Brainerd et al., 2013). Classically, we can avoid an incorrect judgement, by the inclusion of a conditionalizing parameter ($p(A\&B|x) > p(B)$), thereby employing more resources. In QT, we can immediately allow $p(A\&B) > p(B)$, without additional resources. If a cognitive agent lives in a world where she encounters plenty of instances of $p(A\&B) > p(B)$, then it makes sense to employ quantum-like representations, instead of classical ones. This argument has nothing to do with 'physical' advantages in computation in QT and is just one of whether the structure of environmental information matches well the employed representations versus not. A similar analogy can be built in relation to temporal/ memory situations, though precisely how is the topic of future work.

In conclusion, an examination of the TB inequality revealed a novel way to advance our understanding of constructive processes in memory. The report of a violation of the TB inequality with change judgements increases confidence in a claim that quantum-like representations is a plausible way to understand some aspects of human behavior (Manning, 2021).

Chapter Five: General Discussion

Challenging the nature of rationality is at the heart of this work. That is, if something is based on reason, logic and cogency, how could there be alternative interpretations of the same data? Motivated by the notion that CPT has limitations in its explanatory power (see Tversky & Kahneman, 1983), we employed a range of probabilistic models, notably quantum models, to better understand the empirical data in our work. In other words, we explored how different probabilistic frameworks can be used to make varying 'rational' interpretations of data. Specifically, different mathematical axioms (e.g., commutativity, incompatible questions) produce striking implications that cannot be understated. That is, the ability to compare whether classical or quantum properties more closely match observation is a key issue. We now briefly revisit and summarise each chapter before offering some concluding remarks.

5.1 The Bell Bound

In Chapter 2, we designed an experimental paradigm that enabled the application of the Bell Bound in the interaction of two people. Crucially, we employed violations of locality as an information resource, that is, participants had the opportunity to check on their counterpart at will. We created an assortment of PD games that were adjusted to 'bake' the sensitivity to context into the structure of the task. Importantly, there was no consistent optimal strategy for confessing or defecting, rather this was dependent on the payoff matrix presented.

Across five experiments, we explored whether the Bell framework could be broken using variations of the PD game. Experiments 1-3 served to finetune the paradigm and our main analyses focused on Experiments 4 and 5 in the series. We examined how the data compared across both classical and quantum models. We found that participants' S values violated Bell's bound ($S>2$) and so showed they were sensitive to this context. However, recall that the payoff matrices were designed in a way that encouraged checking in some contexts and not others. Indeed, we also allowed participants to check on their counterpart, violating locality by giving them a direct measurement of what their counterpart is doing. Against the backdrop of this experimental setup, our results are not surprising. But there is a more subtle question posed by our work: that is, can choice statistics be modelled in a non-contextual way (that is, in a single table of frequencies)? We employed classical and quantum models to show how they fitted to the empirical data. Our quantum model produced a superior fit, bearing in mind that the empirical data and the quantum model both exceeded the Bell bound to a similar extent. However, our classical model is bounded at S=2 and so produced fits inconsistent with empirical data. It is exactly this subtle point that offers a novel contribution to this field.

In these experiments, we offer a way to understand the notion of supercorrelation in the interaction between agents. All behavioural scientists are familiar with the idea of correlation. How can we have a correlation that is stronger than perfect correlation? The experiments in Chapter 2 primarily aim at developing this insight: given two questions for each of two agents, supercorrelation means perfect correlation for some question pairs and perfect anti-correlation for others. That is, supercorrelation is about perfect coordination in a way that is contextually sensitive to the questions asked by each agent. How relevant is the notion of supercorrelation in behavioural sciences? First, empirically we showed results which do demonstrate that participant behaviour can indicate supercorrelation, which shows that participants are sensitive to supercorrelation structure. Also, we think that it is interesting that in such a case a simple Bayesian model cannot describe choice results (the point has been known from physics, but it is novel in psychology). Second, our work allows a new direction for research, concerning whether supercorrelation situations arise (more) naturally in behaviour.

We would now like to offer our insights for future directions of research. Traditionally, Bell experiments were conducted on particles to refine intuitions in classical physics that physical interactions do not propagate instantly across space. Of course, one of the big challenges we faced was transforming these ideas from physics to psychology or, put differently, from particles to participants. The biggest drawback of this paradigm related to the artificial nature of the interactions. Each participant was partnered with a hypothetical participant (who played stochastically) in a game that offered opportunity to pass information to one another. Indeed, while this is common for PD games, it is nevertheless a highly contrived paradigm that may work in only very selective environments. This is not an issue for physicists, as they have no direct control over either particle, besides the implementation of the measurement tool itself. In terms of future research, experiments could be devised where both players in the PD are real participants in a round-robin style tournament. Moreover, whilst the dynamics of PD games can be seen in everyday life (in terms of trade-offs and response dependencies), paradigms that explore alternative frameworks would be welcomed.

5.2 Constructive Influences

In Chapter 3, we examined constructive influences and whether the expression of an impression changed evaluations made in the future. The idea that an action can change the system is not new, but it does draw parallels with QT. That is, if we were to consider two questions, 1) whether we are happy, and 2) whether we like our job, the order in which we respond to the questions can result in striking differences in how we answer.

In this chapter, we attempted to replicate and extend the work of White et al. (2014). Briefly, we explored whether the expression of an opinion on a positively or negatively valenced advertisement would change evaluations of an advertisement rated later (the EB). We also aimed to establish why different people have different propensities to exhibit an EB. Our approach was to utilise measures of self-reflection, such as mindfulness and metacognition. Measures of self-reflection were employed because it seemed plausible that those who reviewed their evaluations about the advertisements would differ from those who did not.

We conducted a series of experiments, and we examined a major prediction from quantum cognitive models, that a measurement (e.g., a judgment or a decision) can alter the relevant representations in a certain way. Even though the main point has been demonstrated before (with the work of White et al., 2014), the objective here was to consider whether measures of metacognition or empathy might moderate this effect. Such a link between constructive influences and metacognition/ empathy has been motivated in a variety of ways. However, we could not replicate the original effect and so we could not pursue these elaborative hypotheses. Instead, our results offer some promise concerning an order effect of negative vs. positive information in sequential processing: it appears that negative information ‘persists’ more in how subsequent information is assessed, relative to positive information. This is an interesting effect, which relates to various discussions on the primary of positive vs. negative information in cognitive processes (e.g., Sharot et al., 2012).

In this chapter, we offered various lines of reasoning for why this effect occurred, such as negative stimuli attracting more attention than their positive counterparts (Pratto & John, 2001; Field, 2006) or whether greater subjective weight is applied to negative stimuli (Rozin & Royzman, 2001; Hamlin et al., 2010). This is not to say positive EBs do not exist (because there is substantial evidence for their prevalence in society and the literature, Weinstein, 1980; DeJoy, 1989; Sharot, 2011; Fragkaki et al., 2021), but our results showed a trend for negative impressions being more impactful on future judgements. Indeed, these findings might be explained in terms of relevancy to the individual. That is, positive biases described in the literature were more likely to be found when they considered personally relevant information (such as future income or the trajectory of a marriage) and negative biases being reflective of personally irrelevant information (such as the advertisements presented in our paradigm).

With respect to next steps, there are a number of potential routes for developing the experimental paradigm. Firstly, it would be useful developing an up-to-date paradigm that still captures the constructive effects seen in White et al.'s pilot. Specifically, many of the advertisements used in our experiments were outdated (e.g., perceptions of the company Blackberry are likely different in 2021 to what they were in 2014). A related issue is that the valence intensity of the positive and negative stimuli must be similar. In the present work, the positive stimuli were approximately two scale points from the neutral stimuli, whilst the negative stimuli were just over one scale point (see Table 3.2). This creates an issue whereby the positive stimuli are shown to be stronger than their negative counterparts. Thirdly, conceptual misunderstandings of our introspection measures, especially mindfulness, may have resulted in the null findings we observed. That is, our failure to anticipate the implications of different types of mindfulness may have played a pivotal role in explaining the mediation of self-reflection in the EB. Finally, these experiments were conducted during the COVID-19 pandemic and so White's (2014) paradigm had to be amended and designed for online experimentation.

5.3 Temporal Bell

Chapter 4 focused specifically on the TB inequality. Whilst we explored the Bell Bound in Chapter 2 via the interactions between two individuals, TB is more focused on a single person over time. Recall, we used the most canonical version of the TB inequality:

$$C_{12} + C_{23} \leq C_{13} + 1$$

where the sum of the correlations between times points 1,2 and 2,3 are bounded by the correlation between time points 1,3 plus one (analogous to standard Bell Bound of 2). This inequality tests whether there are quantum properties present across three different pairs of time points (that is, t_1, t_2 and t_3). For these quantum properties to be detected accurately, we needed to ensure that at least one of three assumptions were violated: macrorealism, NIM (NSIT/ESIT), and the unidirectional nature of time. Briefly, in the context of human decision making, if macrorealism is violated, this would suggest a person is not definitively in one state or another. Take human memory for example. A violation of macrorealism in a memory experiment would suggest a memory is not stored fully intact and remembering is a reconstructive process that is not set in stone. If the NIM assumption was violated, then this would suggest that the act of measurement has influenced the result. Indeed, if we expected to find constructive influences in decision making, NIM is somewhat of a conundrum. That is, it tasked us to separate 'principled' constructive influences from constructive influences which simply disturbed the system because they were too course. Subsequently, throughout the development of this paradigm, we frequently asked whether, if NIM was violated, was there any evidence that our results still displayed quantum characteristics? Or were our results just the consequence of a crude measurement tool?

The answer to these questions rested on our methodology. We approached the TB inequality with a hypothetical murder trial (see Yearsley & Pothos, 2016), whereby participants undertook the role of a juror and were presented with some evidence across a few days. They were then tasked to determine whether their verdict had changed. Participants were then given an analogous scenario to answer, but instead of answering a 'change' question they were asked for their marginal response at t_1 and then at t_2 (recall, the distinction concerns the question of whether the change question asks whether the verdict had changed from t_1 to t_2, as opposed to requesting a direct verdict). RTs for both marginal and change questions were also recorded to address the concerns of NIM. We wanted to ensure that our 'change' judgements were not the result of two individual judgements and a computation of the difference. That is, if the RTs are different then participants can create the change judgement independently from the individual verdicts.

We conducted five experiments and several pilot and control experiments to finetune the paradigm between iterations. Although the first few experiments were unsuccessful, Experiment 4 violated the TB inequality and Experiment 5 showed the marginal and change

RTs were similar. Violating TB (whilst appeasing the notion of ‘sledgehammer’ NIMs) also indicates a violation of macrorealism. In other words, our results suggested that participants answering a ‘change’ question had no definitive verdict until they were asked the question. The implications of this result are striking. We considered the nature of recollective processes in decision making. While we probably cannot talk about memory processes as such, our paradigm did require participants to recollect various pieces of information in a certain way, in order to answer the questions in the paradigm. The key question here is whether recollection of this kind changes the relevant representations or not. If changes do occur, then we can talk about representations having quantum-like properties. To assess this possibility, we employed the temporal Bell inequalities, finding evidence for corresponding violations. Why does this matter? Because it shows a specific way in which constructive influences in memory more generally might arise, be tested, and be explained. Clearly, as noted, our work cannot talk about memory in general, though we think that our work offers a promising direction in this respect.

Finally, we consider directions for future research. Most pertinent to Chapter 4 is that we conducted numerous iterations of the paradigm before finally violating the TB inequality. This work had plenty of technical challenges and was perhaps the first empirical demonstration of a TB violation in human decision makers. However, one must wonder whether our results can be easily replicated when they were forged under very specific conditions. A more convincing result could be obtained by replicating the current experiment in the real world. In addition to this, future research examining TB in human decision makers should also explore establishing TB violations in other contexts.

5.4 Conclusion

To conclude, we have shown the utility of QT in understanding human decision making. This was primarily illustrated in the Bell framework (Chapters 2 and 4), though it should be noted that the constructive influences research (Chapter 3) was also heavily influenced by QT. The Bell Bound project offered unique insights regarding the interaction between two individuals, namely, sensitivity to context and the propensity of an individual to invoke violations of locality by checking on their counterpart in specific situations and not checking in others. The constructive influences work did not corroborate White et al.’s (2014) seminal research on the EB nor was it successful in explaining its underlying mechanism. Nevertheless, we did find evidence of constructive influences between the first and second advertisements, specifically that negative stimuli rated first had a greater impact on the second positive stimuli than vice versa. Finally, our work on TB served to explore quantum-like structures in human memory. We empirically demonstrated a violation of the TB inequality whilst providing reasonable support for the NIM assumption. Results showed that jurors did not have a definitive verdict, but instead produced one when they were asked.

References

Adenier, G., & Khrennikov, A. Y. (2017). Test of the no-signaling principle in the Hensen loophole-free CHSH experiment. *Fortschritte der Physik, 65*(9), 1-16. https://doi.org/10.1002/prop.201600096

Aerts, D. (1982). Example of a macroscopical classical situation that violates Bell inequalities. *Lettere al Nuovo Cimento (1971-1985), 34*(4), 107-111.

Aerts, D. (1990). An attempt to imagine parts of the reality of the micro-world. In J. Mizerski, A. Posiewnik, J. Pykacz, & M. Zukowski (Eds.), *Problems in Quantum Physics* (pp. 3-25). Singapore: World Scientific.

Aerts, D. (2014). Quantum and concept combination, entangled measurements, and prototype theory. *Topics in Cognitive Science, 6*(1), 129-137. https://doi.org/10.1111/tops.12073

Aerts, D., Sozzo, S., & Veloz, T. (2015). Quantum structure of negation and conjunction in human thought. *Frontiers in Psychology, 6*, 1-30. https://doi.org/10.3389/fpsyg.2015.01447

Aerts, D., Sozzo, S., & Veloz, T. (2016). New fundamental evidence of non-classical structure in the combination of natural concepts. *Philosophical Transactions Royal Society A, 374,* 1-17. https://doi.org/10.1098/rsta.2015.0095

Aerts, S. (2005). *A realistic device that simulates the non-local PR box without communication* [Preprint]. Cornell University. http://arxiv.org/abs/ quant-ph/0504171.

Akhtar, S., Faff, R., Oliver, B., & Subrahmanyam, A. (2011). The power of bad: The negativity bias in Australian consumer sentiment announcements on stock returns. *Journal of Banking & Finance, 35*(5), 1239-1249. https://doi.org/10.1016/j.jbankfin.2010.10.014

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*(1), 126-131. https://doi.org/10.1037/h0027455

Atmanspacher, H., & Filk, T. (2019). Contextuality Revisited: Signaling May Differ From Communicating. In A. de Barros & C. Montemayor (Eds.), *Quanta and Mind – Essays on the Connection between Quantum Mechanics and Consciousness.* https://doi.org/10.1007/978-3-030-21908-6_10

Atmanspacher, H. & Filk, T. (2010). A proposed test of temporal nonlocality in bistable perception. *Journal of Mathematical Psychology, 54*(3), 314-321. https://doi.org/10.1016/j.jmp.2009.12.001

Baer, R. A. (2003). Mindfulness training as a clinical intervention: A conceptual and empirical review. *Clinical Psychology: Science and Practice*, *10*(2), 125-143. https://doi.org/10.1093/clipsy.bpg015

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment, 13*(1), 27-45. https://doi.org/10.1177/1073191105283504

Baer, R. A., Smith, G. T., Lykins, E., Button, D., Krietemeyer, J., Sauer, S., ... & Williams, J. M. G. (2008). Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples. *Assessment*, *15*(3), 329-342. https://doi.org/10.1177/1073191107313003

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders, 34*(2), 163-175.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge University Press.

Basieva, I., Cervantes, V.H., Dzhafarov, E.N., Khrennikov, A. (2019). True contextuality beats direct influences in human decision making. *Journal of Experimental Psychology: General, 148*(11), 1925-1937. https://doi.org/10.1037/xge0000585

Batson, C. D., & Ahmad, N. (2001). Empathy-induced altruism in a prisoner's dilemma II: What if the target of empathy has defected?. *European Journal of Social Psychology, 31*(1), 25-36. https://doi.org/10.1002/ejsp.26

Batson, C. D., & Moran, T. (1999). Empathy-induced altruism in a prisoner's dilemma. *European Journal of Social Psychology, 29*(7), 909-924. https://doi.org/10.1002/(SICI)1099-0992(199911)29:7%3C909::AID-EJSP965%3E3.0.CO;2-L

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323-370. https://doi.org/10.1037/1089-2680.5.4.323

Bell, J. S. (1964). On the Einstein-Podolsky-Rosen paradox. *Physics, 1*(3), 195-200.

Bell, J. S. (1987). *Speakable and unspeakable in quantum mechanics.* Cambridge University Press.

Bergus, G. R., Chapman, G. B., Levy, B. T., Ely, J. W., & Oppliger, R. A. (1998). Clinical diagnosis and order information. *Medical Decision Making, 18*(4), 412-417. https://doi.org/10.1177/0272989X9801800409

Bernstein, D. M., & Loftus, E. F. (2009). How to tell if a particular memory is true or false. *Perspectives on Psychological Science, 4*(4), 370-374. https://doi.org/10.1111/j.1745-6924.2009.01140.x

Blasiak, P., Pothos, E. M., Yearsley, J. M., Gallus, C., & Borsuk, E. (2021). Violations of locality and free choice are equivalent resources in Bell experiments. *Proceedings of the National Academy of Sciences, 118*(17), 1-9. https://doi.org/10.1073/pnas.2020569118

Bless, H., & Schwarz, N. (2010). Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. *Advances in Experimental Social Psychology.* (pp. 319-373). Academic Press. https://doi.org/10.1016/S0065-2601(10)42006-7

Bordalo, P., Gennaioli, N., & Shleifer, A. (2017). Memory, attention, and choice. *The Quarterly Journal of Economics, 135*(3), 1399-1442. https://doi.org/10.1093/qje/qjaa007

Botti, S., Orfali, K., & Iyengar, S. S. (2009). Tragic choices: Autonomy and emotional responses to medical decisions. *Journal of Consumer Research*, *36*(3), 337-352. https://doi.org/10.1086/598969

Brainerd, C. J., Wang, Z., & Reyna, V. F. (2013). Superposition of episodic memories: Overdistribution and quantum models. *Topics in Cognitive Science*, *5*(4), 773-799. https://doi.org/10.1111/tops.12039

Broekaert, J. B., Busemeyer, J. R., & Pothos, E. M. (2020). The disjunction effect in two-stage simulated gambles. An experimental study and comparison of a heuristic logistic, Markov and quantum-like model. *Cognitive Psychology, 117*, 101262. https://doi.org/10.1016/j.cogpsych.2019.101262

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin, 109*(2), 204–223. doi:10.1037/0033-2909.109.2.204

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*(4), 822-848. https://doi.org/10.1037/0022-3514.84.4.822

Brown, K. W., Ryan, R. M., & Creswell, J. D. (2007). Mindfulness: Theoretical foundations and evidence for its salutary effects. *Psychological Inquiry, 18*(4), 211-237. https://doi.org/10.1080/10478400701598298

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition, 5*(1), 73-99. https://doi.org/10.1016/0010-0277(77)90018-X

Bruza, P. D., Kitto, K., Ramm, B., & Sitbon, L. (2015). A probabilistic framework for analysing the compositionality of conceptual combinations. *Journal of Mathematical Psychology, 67*, 26-38. https://doi.org/10.1016/j.jmp.2015.06.002

Bruza, P. D., Kitto, K., Nelson, D., & McEvoy, C. (2009). Is there something quantum-like about the human mental lexicon?. *Journal of Mathematical Psychology, 53*(5), 362-377. https://doi.org/10.1016/j.jmp.2009.04.004

Budroni, C., Fagundes, G., & Kleinmann, M. (2019). Memory cost of temporal correlations. *New Journal of Physics, 21*, 093018. https://doi.org/10.1088/1367-2630/ab3cb4

Busemeyer, J. R., & Bruza, P. D. (2012). *Quantum models of cognition and decision.* Cambridge University Press.

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. (2011). A quantum theoretical explanation for probability judgement errors. *Psychological Review, 118*(2), 193-218. https://doi.org/10.1037/a0022542

Carpenter, J., Sherman, M.T., Kievit, R.A., Seth, A.K., Lau, H. & Fleming, S.M. (2019) Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General, 148*(1), 51-64. https://doi.org/10.1037/xge0000505

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgements of learning. *Psychonomic Bulletin & Review, 14*(1), 107-111.

Cebolla, A., Garcia-Palacios, A., Soler, J., Guillén, V., Baños, R., & Botella, C. (2012). Psychometric properties of the Spanish validation of the Five Facets of Mindfulness Questionnaire (FFMQ). *The European Journal of Psychiatry*, *26*(2), 118-126. https://dx.doi.org/10.4321/S0213-61632012000200005

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287-291. https://doi.org/10.1016/j.tics.2006.05.007

Chater, N., Vlaev, I., & Grinberg, M. (2008). A new consequence of Simpson's paradox: Stable cooperation in one-shot prisoner's dilemma from populations of individualistic learners. *Journal of Experimental Psychology: General, 137*(3), 403-421. https://doi.org/10.1037/0096-3445.137.3.403

Christopher, M. S., Neuser, N. J., Michael, P. G., & Baitmangalkar, A. (2012). Exploring the psychometric properties of the five facet mindfulness questionnaire. *Mindfulness*, *3*(2), 124-131. DOI: 10.1007/s12671-011-0086-x

Clauser, J. F., Horne, M. A., Shimony, A., & Holt, R. A. (1969). Proposed experiment to test local hidden-variable theories. *Physical Review Letters, 23*(15), 880-884.

Cohen, T. R. (2010). Moral emotions and unethical bargaining: The differential effects of empathy and perspective taking in deterring deceitful negotiation. *Journal of Business Ethics, 94*(4), 569-579. https://doi.org/10.1007/s10551-009-0338-z

Conte, E., Khrennikov, A., Todarello, O., & Federici, A. (2008). A preliminary experimental verification on the possibility of Bell inequality violation in mental states. *Neuroquantology, 6*(3), 214-221.

Conway, M. (1994). Flashbulb Memories (1st ed.). Psychology Press. https://doi.org/10.4324/9780203775820

Copen, C. E., Daniels, K., Vespa, J., & Mosher, W. D. (2012). *First Marriages in the United States: Data From the 2006–2010 National Survey of Family Growth*. https://www.cdc.gov/nchs/data/nhsr/nhsr049.pdf

Creswell, J. D., & Lindsay, E. K. (2014). How does mindfulness training affect health? A mindfulness stress buffering account. *Current Directions in Psychological Science*, *23*(6), 401-407.

Dan-Glauser, E. S., & Scherer, K. R. (2011). The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behavior Research Methods, 43*(2), 468-477. https://doi.org/10.3758/s13428-011-0064-1

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10*, 85.

de Bruin, E. I., Topper, M., Muskens, J. G., Bögels, S. M., & Kamphuis, J. H. (2012). Psychometric properties of the Five Facets Mindfulness Questionnaire (FFMQ) in a meditating and a non-meditating sample. *Assessment*, *19*(2), 187-197. https://doi.org/10.1177/1073191112446654

DeJoy, D. M. (1989). The optimism bias and traffic accident risk perception. *Accident Analysis & Prevention*, *21*(4), 333-340. https://doi.org/10.1016/0001-4575(89)90024-9

DeSoto, K. A. (2014). *Confidence ratings in cognitive psychology experiments: Investigating the relationship between confidence and accuracy in memory*. SAGE Publications, Ltd.. https://dx.doi.org/10.4135/978144627305013507683

Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. *Frontiers in Psychology*, 10, 2755. https://doi.org/10.3389/fpsyg.2019.02755

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Sage Publications.

Dzhafarov, E. N., Kujala, J. V., Cervantes, V. H., Zhang, R., & Jones, M. (2016). On contextuality in behavioural data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2068), 20150234.

Eastwood, J. D., Smilek, D., & Merikle, P. M. (2003). Negative facial expression captures attention and disrupts performance. *Perception & Psychophysics, 65*(3), 352-358. https://doi.org/10.3758/BF03194566

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist*, *13*(4), 277-287. https://doi.org/10.1027/1016-9040.13.4.277

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114-38. DOI: 10.1257/mic.3.2.114

El Saadawi, G. M., Azevedo, R., Castine, M., Payne, V., Medvedeva, O., Tseytlin, E., ... & Crowley, R. S. (2010). Factors affecting feeling-of-knowing in a medical intelligent tutoring system: the role of immediate feedback as a metacognitive scaffold. *Advances in Health Sciences Education, 15*(1), 9-30. https://doi.org/10.1007/s10459-009-9162-6

Emary, C. (2017). Ambiguous measurements, signaling, and violations of Leggett-Garg inequalities. *Physical Review A, 96*(4), 042102. https://doi.org/10.1103/PhysRevA.96.042102

Emary, C., Lambert, N., & Nori, F. (2015). Leggett-Garg inequalities. *Reports on Progress in Physics, 77*, 016001. https://doi.org/10.1088/0034-4885/77/1/016001

Epley, N., Caruso, E. M., & Bazerman, M. H. (2006). When perspective taking increases taking: Reactive egoism in social interaction. *Journal of Personality and Social Psychology, 91*(5), 872-889. https://psycnet.apa.org/doi/10.1037/0022-3514.91.5.872

Erisman, S. M., & Roemer, L. (2010). A preliminary investigation of the effects of experimentally induced mindfulness on emotional responding to film clips. *Emotion, 10*(1), 72.

Feldman, G., Hayes, A. F., Kumar, S., Greeson, J., & Laurenceau, J.-P. (2007). Mindfulness and emotion regulation: The development and initial validation of the Cognitive and Affective Mindfulness Scale-Revised (CAMS-R). *Journal of Psychopathology and Behavioral Assessment, 29*(3), 177–190.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 678-693. https://psycnet.apa.org/doi/10.1037/a0014928

Field, A. P. (2006). Watch out for the beast: Fear information and attentional bias in children. *Journal of Clinical Child and Adolescent Psychology, 35*(3), 431-439. https://doi.org/10.1207/s15374424jccp3503_8

Fine, A. (1982). Joint distributions, quantum correlations and commuting observables. *Journal of Mathematical Physics, 23*, 1306-1310. https://doi.org/10.1063/1.525514

Fischhoff, B. (2015). The realities of risk-cost-benefit analysis. *Science*, *350*(6260). https://doi.org/10.1126/science.aaa6516

Fisher, N. R., Mead, B. R., Lattimore, P., & Malinowski, P. (2017). Dispositional mindfulness and reward motivated eating: The role of emotion regulation and mental habit. *Appetite*, *118*, 41-48. https://doi.org/10.1016/j.appet.2017.07.019

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906-911. https://psycnet.apa.org/doi/10.1037/0003-066X.34.10.906

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338-1349. https://doi.org/10.1098/rstb.2011.0417

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Fox, E., Lester, V., Russo, R., Bowles, R. J., Pichler, A., & Dutton, K. (2000). Facial expressions of emotion: Are angry faces detected more efficiently?. *Cognition and Emotion*, *14*(1), 61-92. https://doi.org/10.1080/026999300378996

Fragkaki, I., Maciejewski, D. F., Weijman, E. L., Feltes, J., & Cima, M. (2021). Human responses to Covid-19: The role of optimism bias, perceived severity, and anxiety. *Personality and Individual Differences, 176*, 110781. https://doi.org/10.1016/j.paid.2021.110781

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42. DOI: 10.1257/089533005775196732

Gabbert, F., Memon, A., & Allan, K. (2003). Memory conformity: Can eyewitnesses influence each other's memories for an event?. *Applied Cognitive Psychology, 17*(5), 533-543. https://doi.org/10.1002/acp.885

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin and Review, 10*(4), 843-876.

Garland, E. L., Farb, N. A., R. Goldin, P., & Fredrickson, B. L. (2015). Mindfulness broadens awareness and builds eudaimonic meaning: A process model of mindful positive emotion regulation. *Psychological Inquiry*, *26*(4), 293-314.

Greco, V., & Roger, D. (2001). Coping with uncertainty: The construction and validation of a new measure. *Personality and Individual Differences*, *31*(4), 519-534. https://doi.org/10.1016/S0191-8869(00)00156-2

Greenberg, D. L. (2004). President Bush's false [flashbulb] memory of 9/11/01. *Applied Cognitive Psychology, 18*(3), 363-370. https://doi.org/10.1002/acp.1016

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357-364. https://doi.org/10.1016/j.tics.2010.05.004

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science*, *23*(1), 3-10. https://doi.org/10.1177%2F0963721413498891

Halliwell, J. J. (2014). Two proofs of Fine’s theorem. *Physics Letters A, 378*(40), 2945-2950. https://doi.org/10.1016/j.physleta.2014.08.012

Halliwell, J. J. (2016). Leggett-Garg correlation functions from a noninvasive velocity measurement continuous in time. *Physical Review A, 94*(5), 052114.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450*(7169), 557-559. https://doi.org/10.1038/nature06288

Hamlin, J., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental Science, 13*(6), 923-929. https://doi.org/10.1111/j.1467-7687.2010.00951.x

Hampton, J. A. (1988). Overextension of conjunctive concepts: Evidence for a unitary model for concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 12-32. https://psycnet.apa.org/doi/10.1037/0278-7393.14.1.12

Hansen, C. H., & Hansen, R. D. (1988). Finding the face in the crowd: An anger superiority effect. *Journal of Personality and Social Psychology, 54*(6), 917-924. https://psycnet.apa.org/doi/10.1037/0022-3514.54.6.917

Harris, R. (2009). *ACT made simple: A quick-start guide to ACT basics and beyond.* New Harbinger Publications.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*(4), 208-216. https://psycnet.apa.org/doi/10.1037/h0022263

Haven, E. & Khrennikov, A. (2013). *Quantum Social Science.* Cambridge University Press.

Higham, P. A., & Higham, D. P. (2019). New improved gamma: Enhancing the accuracy of Goodman–Kruskal's gamma using ROC curves. *Behavior Research Methods*, *51*(1), 108-125. https://doi.org/10.3758/s13428-018-1125-5

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 57-80. https://psycnet.apa.org/doi/10.1037/a0013865

Hohaus, L. C., & Spark, J. (2013). Getting better with age: Do mindfulness & psychological well-being improve in old age?. *European Psychiatry*, *28*(S1), 1-1. https://doi.org/10.1016/S0924-9338(13)77295-X

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1-55. https://doi.org/10.1016/0010-0285(92)90002-J

Hölzel, B. K., Lazar, S. W., Gard, T., Schuman-Olivier, Z., Vago, D. R., & Ott, U. (2011). How does mindfulness meditation work? Proposing mechanisms of action from a conceptual and neural perspective. *Perspectives on Psychological Science, 6*(6), 537-559. https://doi.org/10.1177%2F1745691611419671

Hope, L., Ost, J., Gabbert, F., Healey, S., & Lenton, E. (2008). "With a little help from my friends…": The role of co-witness relationship in susceptibility to misinformation. *Acta Psychologica, 127*(2), 476-484. https://doi.org/10.1016/j.actpsy.2007.08.010

Howe, M. L. (2011). The adaptive nature of memory and its illusions. *Current Directions in Psychological Science*, *20*(5), 312-315. https://doi.org/10.1177%2F0963721411416571

Hussain, D. (2015). Meta-cognition in mindfulness: A conceptual analysis. *Psychological Thought, 8*(2), 132-141. https://doi.org/10.23668/psycharchives.1972

Jelicic, M., Smeets, T., Peters, M. J., Candel, I., Horselenberg, R., & Merckelbach, H. (2006). Assassination of a controversial politician: Remembering details from another non-existent film. *Applied Cognitive Psychology, 20*(5), 591-596. https://doi.org/10.1002/acp.1210

Jemstedt, A., Kubik, V., & Jönsson, F. U. (2017). What moderates the accuracy of ease of learning judgements?. *Metacognition and Learning*, *12*(3), 337-355. https://doi.org/10.1007/s11409-017-9172-3

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3-28.

Jones, G. R., & Hill, C. W. (1988). Transaction cost analysis of strategy-structure choice. *Strategic Management Journal*, *9*(2), 159-172. https://doi.org/10.1002/smj.4250090206

Karunamuni, N., & Weerasekera, R. (2019). Theoretical foundations to guide mindfulness meditation: A path to wisdom. *Current Psychology, 38*(3), 627-646. https://doi.org/10.1007/s12144-017-9631-7

Kellen, D., Singmann, H., & Batchelder, W. H. (2018). Classic-Probability Accounts of Mirrored (Quantum-Like) Order Effects in Human Judgements. *Decision, 5*, 323–338.

Khoury, B., Lecomte, T., Fortin, G., Masse, M., Therien, P., Bouchard, V., ... & Hofmann, S. G. (2013). Mindfulness-based therapy: A comprehensive meta-analysis. *Clinical Psychology Review*, *33*(6), 763-771. https://doi.org/10.1016/j.cpr.2013.05.005

Khrennikov, A. (2004). On quantum-like probabilistic structure of mental information. *Open Systems and Information Dynamics, 11*(3), 267-275. https://doi.org/10.1023/B:OPSY.0000047570.68941.9d

Kiken, L. G., Garland, E. L., Bluth, K., Palsson, O. S., & Gaylord, S. A. (2015). From a state to a trait: Trajectories of state mindfulness in meditation during intervention predict changes in trait mindfulness. *Personality and Individual differences*, *81*, 41-46. https://doi.org/10.1016/j.paid.2014.12.044

Kim, H., & Han, H. S. K. (2016). A validation study of the Toronto empathy questionnaire-Korean version. *Korean Journal of Clinical Psychology, 35*(4), 809-821.

Kleinmann, M., Guehne, O., Portillo, J. R., Larsson, J., & Cabello, A. (2011). Memory cost of quantum contextuality. *New Journal of Physics, 13*, 113011. https://doi.org/10.1088/1367-2630/13/11/113011

Kourmousi, N., Amanaki, E., Tzavara, C., Merakou, K., Barbouni, A., & Koutras, V. (2017). The Toronto empathy questionnaire: reliability and validity in a nationwide sample of Greek teachers. *Social Sciences, 6*(2), 62.

Kvam, P. D., Pleskac, T. J., Yu, S., & Busemeyer, J. R. (2015). Interference effects of choice on confidence: quantum characteristics of evidence accumulation. *Proceedings of the National Academy of Sciences, 112*(34), 10645-10650. https://doi.org/10.1073/pnas.1500688112

Lau, M. A., Bishop, S. R., Segal, Z. V., Buis, T., Anderson, N. D., Carlson, L., ... & Devins, G. (2006). The Toronto mindfulness scale: Development and validation. *Journal of Clinical Psychology*, *62*(12), 1445-1467. https://doi.org/10.1002/jclp.20326

Lavner, J. A., Karney, B. R., & Bradbury, T. N. (2013). Newlyweds' optimistic forecasts of their marriage: For better or for worse?. *Journal of Family Psychology, 27*(4), 531-540. https://doi.apa.org/doi/10.1037/a0033423

Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004). Measuring empathy: Reliability and validity of the empathy quotient. *Psychological Medicine, 34*(5), 911-920. https://doi.org/10.1017/S0033291703001624

Leggett, A. J. & Garg, A. (1985). Quantum mechanics versus macroscopic realism: is the flux there when nobody looks? *Physical Review Letters, 54*, 857-860. https://doi.org/10.1103/PhysRevLett.54.857

Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: knowledge partitioning in function learning. *Journal of Experimental Psychology: General, 131*(2), 163-193. https://psycnet.apa.org/doi/10.1037/0096-3445.131.2.163

Liberman, N., Förster, J., & Higgins, E. T. (2007). Completed versus interrupted priming: Reduced accessibility from post-fulfillment inhibition. *Journal of Experimental Social Psychology, 43*(2), 258-264. https://doi.org/10.1016/j.jesp.2006.01.006

Lindsay, E. K., & Creswell, J. D. (2017). Mechanisms of mindfulness training: Monitor and Acceptance Theory (MAT). *Clinical Psychology Review*, *51*, 48-59. https://doi.org/10.1016/j.cpr.2016.10.011

Livingston, J. A. (2003). *Metacognition: An Overview.* Institute of Education Sciences.

Loftus, E. F., & Greenspan, R. L. (2017). If I'm certain, is it true? Accuracy and confidence in eyewitness memory. *Psychological Science in the Public Interest, 18*(1), 1-2. https://doi.org/10.1177%2F1529100617699241

Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about “weapon focus”. *Law and Human Behavior, 11*(1), 55-62.

Lombardi, W. J., Higgins, E. T., & Bargh, J. A. (1987). The role of consciousness in priming effects on categorization: Assimilation versus contrast as a function of awareness of the priming task. *Personality and Social Psychology Bulletin*, *13*(3), 411-429. https://doi.org/10.1177%2F0146167287133009

MacLaverty, S. N., & Hertzog, C. (2009). Do age-related differences in episodic feeling of knowing accuracy depend on the timing of the judgement?. *Memory, 17*(8), 860-873. https://doi.org/10.1080/09658210903374537

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: meta-d′, response-specific meta-d′, and the unequal variance SDT model. In *The Cognitive Neuroscience of Metacognition* (pp. 25-66). DOI: 10.1007/978-3-642-45190-4_3

Manning, J. R. (2021). Episodic memory: Mental time travel or a quantum “memory wave” function?. *Psychological review*, *128*(4), 711-725. https://psycnet.apa.org/doi/10.1037/rev0000283

Martin, L. L. (1986). Set/reset: Use and disuse of concepts in impression formation. *Journal of Personality and Social Psychology, 51*(3), 493-504.

Martin, L. L., & Shirk, S. (2007). Set/Reset and Self-Regulation: Do Contrast Processes Play a Role in the Breakdown of Self-Control? In D. A. Stapel & J. Suls (Eds.), *Assimilation and Contrast in Social Psychology.* (pp. 207–225). Psychology Press.

McKenzie, C. R. M., Lee, S. M., & Chen, K. K. (2002). When negative evidence increases confidence: Change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making, 15*(1), 1–18. https://doi.org/10.1002/bdm.400

Mogg, K., Bradley, B. P., & Williams, R. (1995). Attentional bias in anxiety and depression: The role of awareness. *British Journal of Clinical Psychology, 34*(1), 17-36. https://doi.org/10.1111/j.2044-8260.1995.tb01434.x

Moore, D. W. (2002). Measuring new types of question-order effects: Additive and subtractive. *The Public Opinion Quarterly, 66*(1), 80-91. https://www.jstor.org/stable/3078697

Nelson, A., & Shiffrin, R. (2010). SARKAE - Modeling the Co-Evolution of Event Memory and Knowledge. *Proceedings of the Annual Meeting of the Cognitive Science Society, 32*, 254-259. https://escholarship.org/uc/item/4fk0b6jq

Nelson, D. L., & McEvoy, C. (2007). Entangled Associative Structures and Context. In *AAAI Spring Symposium: Quantum Interaction* (pp. 98-105). https://www.aaai.org/Papers/Symposia/Spring/2007/SS-07-08/SS07-08-015.pdf

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109-133. https://psycnet.apa.org/doi/10.1037/0033-2909.95.1.109

Nelson, T. O., & Dunlosky, J. (1991). When people's judgements of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2*, 267–270.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition?. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing*, (pp. 1-25). Retrieved from shorturl.at/pvyNU

Norman, E., Pfuhl, G., Sæle, R. G., Svartdal, F., Låg, T., & Dahl, T. I. (2019). Metacognition in psychology. *Review of General Psychology*, *23*(4), 403-424. https://doi.org/10.1177%2F1089268019883821

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, *32*(1), 69-84. doi:10.1017/S0140525X09000284

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: a threat advantage with schematic stimuli. *Journal of Personality and Social Psychology, 80*(3), 381. https://psycnet.apa.org/doi/10.1037/0022-3514.80.3.381

ONS. (2019). *Divorces in England and Wales Divorces in England and Wales: 2019*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/divorce/bulletins/divorcesinenglandandwales/2019

Osherson, D. N. & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35-58. https://doi.org/10.1016/0010-0277(81)90013-5

Ost, J., Granhag, P. A., Udell, J., & Roos af Hjelmsäter, E. (2008). Familiarity breeds distortion: The effects of media exposure on false reports concerning media coverage of the terrorist attacks in London on 7 July 2005. *Memory, 16*(1), 76-85. https://doi.org/10.1080/09658210701723323

Ost, J., Vrij, A., Costall, A., & Bull, R. (2002). Crashing memories and reality monitoring: Distinguishing between perceptions, imaginations and 'false memories'. *Applied Cognitive Psychology, 16*(2), 125-134. https://doi.org/10.1002/acp.779

Paterson, H. M., Kemp, R. I., & Forgas, J. P. (2009). Co-witnesses, confederates, and conformity: Effects of discussion and delay on eyewitness memory. *Psychiatry, Psychology and Law*, *16*(sup1), S112-S124. https://doi.org/10.1080/13218710802620380

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition?. *Behavior Research Methods*, *48*(1), 341-348.

Popescu, S., and Rohrlich, D. (1994). Quantum Nonlocality as an Axiom. *Foundations of Physics, 24*, 379-385. https://doi.org/10.1007/BF02058098

Pothos, E. M. & Busemeyer, J. R. (2009). A quantum probability explanation for violations of 'rational' decision theory. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1665), 2171-2178. https://doi.org/10.1098/rspb.2009.0121

Pothos, E. M., Busemeyer, J. R., & Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review*, *120*(3), 679-696. https://psycnet.apa.org/doi/10.1037/a0033142

Pothos, E. M., Lewandowsky, S., Basieva, I., Barque-Duran, A., Tapper, K., & Khrennikov, A. (2021). Information overload for (bounded) rational agents. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1944), 20202957. https://doi.org/10.1098/rspb.2020.2957

Pratto, F., & John, O. P. (1991). Automatic vigilance: the attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*(3), 380-391. https://psycnet.apa.org/doi/10.1037/0022-3514.61.3.380

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453-469. https://doi.org/10.1002/bdm.1883

Reyna, V. F. & Brainerd, C. J. (1998). Fuzzy-trace theory and false memory: new frontiers. *Journal of Experimental Child Psychology*, *71*(2), 194-209. https://doi.org/10.1006/jecp.1998.2472

Reyna, V. F. (2008). A theory of medical decision making and health: fuzzy trace theory. *Medical Decision Making*, *28*(6), 850-865. https://doi.org/10.1177%2F0272989X08327066

Rhodes, M. G. (2016). Judgements of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 65–80). Oxford University Press.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgements of learning (JOLs) on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*(1), 131–148. https://psycnet.apa.org/doi/10.1037/a0021705

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296-320. https://doi.org/10.1207%2FS15327957PSPR0504_2

Schacter, D. L., Guerin, S. A., & St. Jacques, P. L. (2011). Memory distortion: an adaptive perspective. *Trends in Cognitive Sciences, 15*(10), 467-474. https://doi.org/10.1016/j.tics.2011.08.004

Schnyer, D. M., Verfaellie, M., Alexander, M. P., LaFleche, G., Nicholls, L., & Kaszniak, A. W. (2004). A role for right medial prefrontal cortex in accurate feeling-of-knowing judgements: evidence from patients with lesions to frontal cortex. *Neuropsychologia, 42*(7), 957-966. https://doi.org/10.1016/j.neuropsychologia.2003.11.020

Schwartz, B. L. (1994). Sources of information in metamemory: Judgements of learning and feelings of knowing. *Psychonomic Bulletin & Review*, *1*(3), 357-375. https://doi.org/10.3758/BF03213977

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition, 25*(5), 638-656. https://doi.org/10.1521/soco.2007.25.5.638

Schwarz, N., & Bless, H. (1992). Assimilation and contrast effects in attitude measurement: An inclusion/exclusion model. *Advances in Consumer Research, 19*, 72-77. https://www.acrwebsite.org/volumes/7271/volumes/v19/

Seaward, H. G., & Kemp, S. (2000). Optimism bias and student debt. *New Zealand Journal of Psychology*, *29*(1), 17-19.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology, 24*(4), 449-474. https://doi.org/10.1016/0010-0285(92)90015-T

Sharot, T. (2011). The optimism bias. *Current Biology*, *21*(23), R941-R945. https://doi.org/10.1016/j.cub.2011.10.030

Sharot, T., Kanai, R., Marston, D., Korn, C. W., Rees, G., & Dolan, R. J. (2012). Selectively altering belief formation in the human brain. *Proceedings of the National Academy of Sciences*, *109*(42), 17058-17062. https://doi.org/10.1073/pnas.1205828109

Sharot, T., Velasquez, C. M., & Dolan, R. J. (2010). Do decisions shape preference? Evidence from blind choice. *Psychological Science*, *21*(9), 1231-1235. https://doi.org/10.1177%2F0956797610379235

Sheldon, K. M., Ryan, R., & Reis, H. T. (1996). What makes for a good day? Competence and autonomy in the day and in the person. *Personality and Social Psychology Bulletin*, *22*(12), 1270-1279. https://doi.org/10.1177%2F01461672962212007

Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology, 39*(2), 211–221. https://psycnet.apa.org/doi/10.1037/0022-3514.39.2.211

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750-756. https://doi.org/10.1038/s41562-018-0425-1

Sjödén, B., Granhag, P. A., Ost, J., & Roos Af Hjelmsäter, E. M. M. A. (2009). Is the truth in the details? Extended narratives help distinguishing false "memories" from false "reports". *Scandinavian Journal of Psychology, 50*(3), 203-210. https://doi.org/10.1111/j.1467-9450.2008.00694.x

Smeets, T., Telgen, S., Ost, J., Jelicic, M., & Merckelbach, H. (2009). What's behind crashing memories? Plausibility, belief and memory in reports of having seen non-existent images. *Applied Cognitive Psychology, 23*(9), 1333-1341. https://doi.org/10.1002/acp.1544

Smith, N. K., Cacioppo, J. T., Larsen, J. T., & Chartrand, T. L. (2003). May I have your attention, please: Electrocortical responses to positive and negative stimuli. *Neuropsychologia*, *41*(2), 171-183. https://doi.org/10.1016/S0028-3932(02)00147-1

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: a meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*(3), 315. https://psycnet.apa.org/doi/10.1037/0033-2909.118.3.315

Spreng, R. N., McKinnon, M. C., Mar, R. A., & Levine, B. (2009). The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment, 91*(1), 62-71. https://doi.org/10.1080/00223890802484381

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149. https://doi.org/10.3758/BF03207704

Stankov, L., Kleitman, S., & Jackson, S. A. (2015). Measures of the trait of confidence. In G. Boyle, D. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs.* (pp. 158-189). Academic Press. https://doi.org/10.1016/B978-0-12-386915-9.00007-3

Stieger, S., & Reips, U. D. (2016). A limitation of the Cognitive Reflection Test: Familiarity. *PeerJ*, *4*, e2395. https://doi.org/10.7717/peerj.2395

Storms, G., de Boeck, P., Hampton, J.A., & van Mechelen, I. (1999). Predicting conjunction typicalities by component typicalities. *Psychonomic Bulletin and Review, 6*, 677-684. https://doi.org/10.3758/BF03212978

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers?. *Acta Psychologica, 47*(2), 143-148. https://doi.org/10.1016/0001-6918(81)90005-6

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgements of retention (JORs). *Quarterly Journal of Experimental Psychology, 65*(7), 1376-1396. https://doi.org/10.1080%2F17470218.2012.656665

Tenenbaum, J. B., & Griffiths, T. (2001). Structure learning in human causal induction. In *Advances in Neural Information Processing Systems: Proceedings of the 13th International Conference on Neural Information Processing Systems*. (pp. 52–58). https://papers.nips.cc/paper/1845-structure-learning-in-human-causal-induction.pdf

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. (2011). How to grow a mind: statistics, structure, and abstraction. *Science, 331*(6022), 1279-1285. https://doi.org/10.1126/science.1192788

Toner, B. F. & Bacon, D. (2003). Communication cost of simulating Bell correlations. *Physical Review Letters, 91*(18), 187904. https://doi.org/10.1103/PhysRevLett.91.187904

Totan, T., Dogan, T., & Sapmaz, F. (2012). The Toronto Empathy Questionnaire: Evaluation of Psychometric Properties among Turkish University Students. *Eurasian Journal of Educational Research, 46*, 179-198.

Trueblood, J. S., & Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science*, *35*(8), 1518-1552. https://doi.org/10.1111/j.1551-6709.2011.01197.x

Trueblood, J., & Dasari, A. (2017). The Impact of Presentation Order on the Attraction Effect in Decision-making. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 3374–3379). Cognitive Science Society.

Trueblood, J. S., & Hemmer, P. (2017). The generalized quantum episodic memory model. *Cognitive Science*, *41*(8), 2089-2125. https://doi.org/10.1111/cogs.12460

Tsirelson, B. S. (1980). Quantum generalizations of Bell's inequality. *Letters of Mathematical Physics, 4*, 93–100. https://doi.org/10.1007/BF00417500

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, *90*(4), 293-315. https://psycnet.apa.org/doi/10.1037/0033-295X.90.4.293

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 106*(4), 1039-1061. https://doi.org/10.2307/2937956

Vlaev, I., & Chater, N. (2006). Game relativity: How context influences strategic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 131-149. https://psycnet.apa.org/doi/10.1037/0278-7393.32.1.131

Waddup, O., Blasiak, P., Yearsley, J. M., Wojciechowski, B. W., & Pothos, E. M. (2021). Sensitivity to Context in Human Interactions. *Mathematics, 9*(21), 2784. doi:10.3390/math9212784

Walach, H., Buchheld, N., Buttenmüller, V., Kleinknecht, N., & Schmidt, S. (2006). Measuring mindfulness—the Freiburg mindfulness inventory (FMI). *Personality and Individual Differences*, *40*(8), 1543-1555. https://doi.org/10.1016/j.paid.2005.11.025

Wang, K., Emary, C., Xu, M., Zhan, X., Bian, Z., Xioa, L., & Xue, P. (2018). Violations of a Leggett-Garg inequality without signaling for a photonic qutrit probed with ambiguous measurements. *Physical Review A, 97*(2), 020101. https://link.aps.org/doi/10.1103/PhysRevA.97.020101

Wang, Z., Solloway, T., Shiffrin, R. M., & Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgements. *Proceedings of the National Academy of Sciences, 111*(26), 9431–9436. https://doi.org/10.1073/pnas.1407756111

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*(5), 806-820. https://psycnet.apa.org/doi/10.1037/0022-3514.39.5.806

White, L. C., Barqué-Duran, A., & Pothos, E. M. (2016). An investigation of a quantum probability model for the constructive effect of affective evaluation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2058), 20150142. https://doi.org/10.1098/rsta.2015.0142

White, L. C., Pothos E. M., & Jarrett, M. (2020). The cost of asking: how evaluations bias subsequent judgements. *Decision, 7*(4), 259-286. https://psycnet.apa.org/doi/10.1037/dec0000136

White, L. C., Pothos, E. M., & Busemeyer, J. R. (2014). Sometimes it does hurt to ask: the constructive role of articulating impressions. *Cognition, 133*(1), 48-64. https://doi.org/10.1016/j.cognition.2014.05.015

White, L., Pothos, E. M., & Jarrett, M. (2019). Evidence for constructive influences from simple evaluations. *Cognitive Science Mind Modeling*, 1-6. https://osf.io/pw67e/download

Wilde, M. M. & Mizel, A. (2012). Addressing the clumsiness loophole in a Leggett-Garg test of macrorealism. *Foundations of Physics, 42*, 256-265. https://doi.org/10.1007/s10701-011-9598-4

Wilding, S., Conner, M., Sandberg, T., Prestwich, A., Lawton, R., Wood, C., ... & Sheeran, P. (2016). The question-behaviour effect: A theoretical and methodological review and meta-analysis. *European Review of Social Psychology*, *27*(1), 196-230. https://doi.org/10.1080/10463283.2016.1245940

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10-65. https://doi.org/10.1177%2F1529100616686966

Yearsley, J. M. & Pothos, E. M. (2014). Challenging the classical notion of time in cognition: a quantum perspective. *Proceedings of the Royal Society B: Biological Sciences, 281*(1781), 1471-1479. https://doi.org/10.1098/rspb.2013.3056

Yearsley, J. M. & Pothos, E. M. (2016). Zeno's paradox in decision making. *Proceedings of the Royal Society B: Biological Sciences, 283*(1828), 20160291. https://doi.org/10.1098/rspb.2016.0291

Yearsley, J. M. & Trueblood, J. S. (2018). A Quantum theory account of order effects and conjunction fallacies in political judgements. *Psychonomic Bulletin & Review, 25*, 1517–1525. https://doi.org/10.3758/s13423-017-1371-z

CITY UNIVERSITY LONDON

STATEMENT OF CO-AUTHORS of JOINT PUBLICATIONS

TO WHOM IT MAY CONCERN

Title of publication: Sensitivity to context in human interactions

Name of candidate: Oliver Waddup

Title of research thesis: A Quantum Approach to Human Decision Making

Name of first supervisor: Emmanuel Pothos

We, the undersigned, co-authors of the above publication, confirm that the above publication has not been submitted as evidence for which a degree or other qualification has already been awarded.

We, the undersigned, further indicate the candidate’s contribution to the publication in our joint statement below.

Signature: EMMANUEL POTHOS

Name: EMMANUEL POTHOS
Date: 16.06.2022

Signature: JAMES YEARSLEY

Name: JAMES YEARSLEY
Date: 16.06.2022

Signature: PAWEL BLASIAK

Name: PAWEL BLASIAK
Date: 16.06.2022

Signature: BARTOSZ WOJCIECHOWSKI

Name: BARTOSZ WOJCIECHOWSKI
Date: 16.06.2022

Statement indicating the candidate’s contribution to the publication
{Statement in support of candidate’s contribution to the publication}

I carried out all empirical work (BW laid the foundations for the use of empathy in the bell framework), helped with the computational fits (alongside EP and JY), and had a more minor involvement with the mathematical part (mostly EP, JY and PB).