Understanding Unpredictability:

On the Responses to and the Valence of the Unpredicted and the Unpredictable

Inauguraldissertation

zur

Erlangung des Doktorgrades

der Humanwissenschaftlichen Fakultät

der Universität zu Köln

nach der Promotionsordnung vom 18.12.2018

vorgelegt von

Judith Gerten

aus

Köln

Tag der Abgabe: 23.07.2020

Diese Dissertation wurde von der Humanwissenschaftlichen Fakultät der Universität zu Köln im Dezember 2020 angenommen.

Erstgutachter: Prof. Dr. Sascha Topolinski

Zweitgutachter: Jun.-Prof. Dr. Oliver Genschow

Tag der Disputation: 11.12.2020

Erklärung

Kapitel 2 beruht auf folgendem Manuskript:

Gerten, J., & Topolinski, S. (2019). Shades of surprise: Assessing surprise as a function of degree of deviance and expectation constraints. *Cognition, 192*, 103986. https://doi.org/10.1016/j.cognition.2019.05.023

Beide Autoren waren an der Entwicklung der Idee beteiligt. Ich habe die Datenerhebung überwacht, die Datenanalyse ausgeführt und das Manuskript geschrieben. Der zweite Autor hat zu jedem dieser Schritte wertvolle Vorschläge beigetragen.

Kapitel 3 beruht auf folgendem Manuskript:

Gerten, J., Zürn, M., & Topolinski, S. (2020). The price of predictability – Estimating inconsistency premiums in social interactions. *Manuscript submitted for publication.*

Alle Autoren waren an der Entwicklung der Idee beteiligt. Ich und der zweite Autor haben die Datenerhebung überwacht und die Datenanalyse ausgeführt. Ich habe das Manuskript geschrieben, das durch wertvolle Beiträge des zweiten und dritten Autors profitiert hat.

Judith Gerten, Köln, 23.07.2020

Abstract

Unpredictability constitutes a deeply ingrained phenomenon of our everyday lives. At some times, things happen unpredictedly, breaking our hitherto existing expectations and filling us with surprise; at other times, we neither hold expectations nor make predictions, and what will happen will do so unpredictably. Both the unpredicted and the unpredictable have been subject to extensive previous research endeavors which have spawned a bunch of heterogeneous theories and evidence, controversial debates, and a range of open questions. In this dissertation, I investigate the cause and structure of responses to surprise and the valence of the unpredicted and the unpredictable to foster a successive integration of single threads and to increase the psychological understanding of unpredictability. Chapter 1 introduces the relevant theoretical background and provides an overview on the current literature. Chapter 2 investigates the effects of the degree of deviance and expectation constraints on the behavioral, affective, experiential, and cognitive responses to unpredicted, surprising events. The evidence obtained in two experiments suggests that the key driving mechanism of surprise is unexpectedness and not the ease of making sense of an event. Beyond that, the behavioral, experiential, and cognitive responses to surprise apparently unfold in a dichotomous way, distinguishing between deviance and non-deviance without being sensitive to finer gradations. On the affective dimension, the evidence points towards surprise being inherently valence-free. Chapter 3 transfers the economic principles of risk-return trade-off and risk premium to the psychological domain, investigating whether and what value people attach to predictable social interactions. Across seven experiments, I demonstrate that people are willing to forgo substantial parts of their potential returns to ensure interacting with a predictable (vs. unpredictable) partner. This suggests an overall negative valence of the unpredictable. Chapter 5 concludes with discussing implications, limitations, and future directions of the research presented.

Key words: unpredictability, surprise, expectation, sense-making, uncertainty

Deutsche Kurzzusammenfassung

Unvorhersagbarkeit ist tief in unserem Alltagsleben verankert. Manche Dinge ereignen sich unvorhergesagt, entgegen unseren bislang bestehenden Erwartungen und versetzen uns in Überraschung; anderem hingegen begegnen wir vollkommen erwartungs- und vorhersagelos und was auch immer sich ereignen wird, wird unvorhersagbar sein. Sowohl das Unvorhergesagte als auch das Unvorhersagbare bilden Gegenstand umfangreicher bisheriger Forschungsunterfangen, die eine breite Palette heterogener Theorien und Evidenz, kontroverse Debatten und eine lange Liste offener Fragen hervorgebracht haben. Die vorliegende Dissertation widmet sich der Untersuchung der kausalen Faktoren und Struktur von Reaktionen auf Überraschung sowie der Valenz des Unvorhergesagten und Unvorhersagbaren, um bisherige Einzelfäden sukzessive zusammenzuführen und das psychologische Gesamtverständnis von Unvorhersagbarkeit zu erhöhen. Kapitel 1 beinhaltet eine Einführung in den relevanten theoretischen Kontext und gibt einen Überblick über die bestehende Literaturlandschaft. Kapitel 2 untersucht die Effekte des Devianzgrads und der Restriktivität von Erwartungen auf die behavioralen, affektiven, experientiellen und kognitiven Reaktionen auf Überraschung. Befunde aus zwei Experimenten suggerieren, dass Unerwartetheit der treibende kausale Mechanismus für Überraschung ist und die Leichtigkeit, ein Ereignis in einen Sinnzusammenhang einzubetten, keine zentrale Rolle spielt. Zudem manifestieren sich die behavioralen, experientiellen und kognitiven Überraschungsreaktionen mit einer Differenzierung zwischen devianten und non-devianten Ereignissen auf dichotome Art, zeigen jedoch keine Sensitivität für verschieden starke Devianzgrade. Auf affektiver Dimension implizieren die beobachteten Reaktionsmuster eine inhärente Valenzfreiheit von Überraschung. Kapitel 3 überträgt die ökonomischen Konzepte des Risiko-Ertrags-Verhältnisses und der Risikoprämie auf psychologische Bereiche und untersucht, ob bzw. welchen Wert Personen Vorhersagbarkeit in sozialen Interaktionen zuschreiben. Evidenz aus insgesamt sieben Experimenten verweist auf die Bereitschaft, auf einen substanziellen Teil möglicher

Interaktionsprofite zu verzichten, um eine Interaktion mit einem vorhersagbaren (vs. unvorhersagbaren) Gegenüber sicherzustellen. Dies impliziert eine negative Valenz des Unvorhersagbaren. Kapitel 5 schließt mit einer Diskussion von Implikationen, Limitationen und künftigen Forschungsrichtungen.

Schlagwörter: Unvorhersagbarkeit, Überraschung, Erwartung, Sinn-Findung, Unsicherheit

Danksagung

An dieser Stelle möchte ich meinen Dank an die Personen richten, ohne deren Unterstützung die vorliegende Dissertation so nicht möglich gewesen wäre.

Mein Dank gilt zunächst meinem Erstbetreuer Prof. Dr. Sascha Topolinski. Danke, lieber Sascha, für Deine Wegbegleitung, Deine konstruktiven Impulse und Deinen Ideenreichtum, der mich immer wieder inspiriert und bereichert hat. Ebenso danke ich Jun.-Prof. Dr. Oliver Genschow für seine Unterstützung als Zweitgutachter.

Meinen Kolleginnen und Kollegen möchte ich für die Vielzahl an Gesprächen und Anregungen danken. Besonderer Dank sei an dieser Stelle an unsere studentischen Hilfskräfte gerichtet, die durch Ihre Unterstützung in der Datenerhebung substanziell zu dieser Arbeit beigetragen haben.

Liebe Familie und Freunde, lieber Daniel, ich danke Euch für Euer Dasein. Danke, dass Ihr – auf welchem Weg mit welchem Ziel auch immer – Halt und Orientierung gebt.

Table of Contents

Chapter 1 – Introduction

Life is no linear journey and no marble run on which the course and destination of the ball are foreseeable from the moment we let go of it. Rather, life is a departure into the chaotic wild. At some time, things go according to plan, and at other time, life overwhelms us with its very own plans. And amidst our modern, vibrant, and constantly changing worlds, the only thing we can predict about the future is possibly that it will be largely unpredictable, and the unforeseen will constitute an essential part of our day-to-day lives.

Rising from this science-fictional sounding reality, the focus of the current dissertation branches into two major streams: the unpredicted and the unpredictable. Undoubtedly, we all hold our spontaneous associations with these terms – be it the classic unexpected and embarrassing surprise birthday party, the vehemence with which the corona pandemic hit the world at the beginning of 2020, or the arc of suspense of a Stephen King movie. But let's take a more scientific perspective. Studying the unpredicted and the unpredictable from an information-theoretical map reveals that one stream actually flows into the other, as the amount of unpredictability results from the expected amount of unpredicted events during an event sampling episode (e.g., Schiffer, Ahlheim, Wurm, & Schubotz, 2012; Strange, Duggins, Penny, Dolan, & Friston, 2005). If we have, for instance, a biased coin that always lands on tail, we do not expect any unpredicted events, no matter how often we throw it, and thus there is full predictability. If, in turn, our coin is unloaded, we will expect on average half of all tosses to come up contrary to our predictions, and thus the amount of unpredictability is substantially higher.

Despite this close computational linkage, a distinction between the unpredicted and the unpredictable proves essential from a psychological perspective. The first reason for this derives from their differential temporal orientation. Whereas the unpredicted encompasses events that have already happened and hence takes a retrospective, the unpredictable is

prospectively directed by pointing to future events. Relating thereto, the unpredicted differs from the unpredictable in terms of the epistemic certainty one holds about the accuracy of one's prediction for a specific event. While the temporal completion of an unpredicted event provides post-hoc certainty that one's prediction does not hold true, the temporal open-endedness of an unpredictable future event triggers a-priori uncertainty about the accuracy of one's prediction.

Taking a multi-perspective, the current dissertation investigates the psychological responses to both the unpredicted and the unpredictable. Within this scope, my work intends to answer three central questions. Firstly, what are the key driving determinants of responses to the unpredicted? Chapter 2 probes this question by examining whether surprise responses are mainly driven by unexpectedness or by the ease of making sense of an event. Secondly, do all unpredicted events trigger the same uniform response patterns? To answer this issue, Chapter 2 pitches a dichotomous all-or-nothing account against the assumption of continuous grades of surprise. Thirdly, what is the valence of unpredictability? Chapter 2 examines the affective responses to the unpredicted. Uniting theories from financial economics and psychology, Chapter 3 extends this evidence by investigating the valence of the unpredictable by assessing whether and what value people are willing to forgo to ensure predictability in social interactions.

In the remainder of Chapter 1, I create the conceptual foundations for the subsequently presented research. Specifically, I give an overview of the predominant theories on the causes of surprise and examine the current state of knowledge on whether responses to unpredicted, surprising events unfold in a dichotomous way or are gradually structured. Following this, I provide a summary of the diverging perspectives on the valence of the unpredicted and the unpredictable. Chapter 1 finally closes with an outline of the current research.

1.1 On the Cause and Structure of Surprise

Life is full of unpredicted events that run against our expectations – for example, meeting your former prime school teacher during your vacancy at the other end of the world, learning that fire does not have a shadow, or seeing a cute puppy photograph after a sequence of artificial non-words presented on the screen in a psychological experiment. Commonly, such events are known as *surprise*.

A plethora of previous research has shown that surprise triggers a complex reaction cascade of multiple components (for an overview, see Reisenzein, 2000a). These comprise the cognitive appraisal of an event as unexpected (e.g., Meyer, Niepel, Rudolph, & Schützwohl, 1991; Meyer, Reisenzein, & Schützwohl, 1997), a qualitative feeling of surprise (Reisenzein, 2000b), the disruption of ongoing cognitive and motor activity (e.g., Horstmann, 2005; Scherer, Zentner, & Stern, 2004), the reallocation of attentional resources (see Horstmann, 2015), a range of physiological and neural responses (e.g. Donchin, 1981; Wessel, Jenkinson, Brittain, Voets, Aziz, & Aron, 2016), and expressive manifestations such as verbal exclamations (Reisenzein, Bördgen, Holtbernd, & Matz, 2006) and a distinct facial display (see Reisenzein, Studtmann, & Horstmann, 2013). The debate on surprise-specific affective response components has joined the research agenda only recently, but already succeeded in splitting scientists into a fistful of different positions which either argue that surprise is inherently positive or inherently negative or devoid of any innate valence (for a recent overview, see Reisenzein, Horstmann, & Schützwohl, 2019; see also Chapter 1.2 for a more detailed elaboration).

While there is relatively wide agreement on the measurable outcomes of surprise (leaving aside the heated discussion on the affective component), scientific positions diverge much more on the question of what *causes* a surprise (see also Munnich, Foster, & Keane, 2019; Reisenzein et al., 2019). The current bunch of psychological theories on surprise can be largely

divided into two main approaches: the one focusing on unexpectedness, the other one on sense-making and comprehension. According to the advocates of the *unexpectedness approach*, surprise arises from the disconfirmation of expectations (e.g., Meyer et al., 1997; for an overview, see also Reisenzein et al., 2019). Imagine your brain as an ambitioned control freak that continuously monitors the congruency between your currently activated expectation about a given situation and the incoming information – as long as these two have a match, nothing happens. But as soon as your brain detects significant discrepancies between what you expected to happen and what actually happened, it spreads an organism-wide surprise signal which triggers the above-mentioned response cascade (e.g., Meyer et al., 1997; see also Reisenzein, 2000a). Note that although theoretically referring to any kind of implicit, explicit, active or passive expectation, belief, and mental model (e.g., Macedo & Cardoso, 2019), the vast majority of research from this approach has induced expectations via repetition-based learning (for a comprehensive summary, see Reisenzein et al., 2019), accordingly defining "unexpected" as "an unannounced deviation from the previous mode of presentation" (Meyer et al., 1997, p. 257). The present terminological understanding and later experimental operationalization of "unexpected" will thus refer to this specific understanding.

Advocates of a *sense-making approach*, in turn, argue that surprise reflects the meta-cognitive difficulty of integrating an event into the realm of pre-existing knowledge structures (e.g., Foster & Keane, 2015; Maguire, Maguire, & Keane, 2011). This time, imagine your brain as a little Sherlock Holmes[1] who steadily strives to provide coherent, logical explanations. As long as events can be easily explained, nothing happens. But as soon as there is no preformed explanation ready at hand, the existing knowledge frameworks break down, thereby demanding a restructuration of mental representations until the event makes sense again.

[1] The reference to a miniature creature ("Sherlock Holmes") in the human head is used only metaphorically for stylistic reasons and does not build on a broader homunculus theory.

Despite their divergent perspectives on the causal antecedents of surprise, both unexpectedness and sense-making approaches assume a graded-ness of the surprise responses – the stronger an event disconfirms an expectation or the harder it is to explain, the higher is the resulting level of surprise (see Foster & Keane, 2015; Reisenzein et al., 2006; Teigen & Keren, 2003; but cf., Ludden, Schifferstein, & Hekkert, 2012). The findings that substantiate this claim empirically remain, however, scarce, methodically inconsistent, or merely correlational, and a systematic investigation is still lacking (for an overview, see Reisenzein et al., 2019).

Thus, the overall range of empirical evidence that has shed light on the determinants and response structure of surprise appears fragmentary and rather selective than exhaustive. A unifying approach that merges all threads into a comprehensive big picture is so far missing, paving way for further research. Aiming at addressing this lacuna and at fostering a better understanding of the operating principles of surprise, the work presented in Chapter 2 focuses on the key driving determinants and the structural graded-ness of the behavioral, affective, experiential and cognitive responses to unpredicted, surprising events.

1.2 On the Valence of the Unpredicted and the Unpredictable

If I asked you to remember the last time you encountered something unpredicted or unpredictable and to describe how you felt at that moment, I would probably gather a plethora of responses. Coming back to the introducing examples from Chapter 1.1, you would presumably experience positive affect when seeing a cute puppy photograph after a sequence of artificial non-words presented on the screen, whereas meeting your former prime school teacher in vacancy might have a rather negative valence (unless you were the preferred model pupil). In the same vein, you might enjoy the unpredictable state of not knowing what your partner will give you as a birthday present, whereas unpredictable developments of the stock market likely feel highly aversive. Aiming at providing an integrative overview on hitherto

existing perspectives, the following sections will outline current psychological and neurophysiological theories and findings on the valence of the unpredicted and the unpredictable.

1.2.1 A Psychological Perspective

Although researchers agree that it feels like something to encounter the unpredicted and to be surprised (see Reisenzein, 2000b), there is dissens on *how* it feels. Does surprise trigger positive affect? Is it affectively negative? Or doesn't it feel in a particular way at all but is just the cognitive diagnosis of unexpectedness or difficulty of sense-making?

From a psychological perspective, the first option – that surprise evokes exclusively positive affect – receives only sparse empirical support (Fontaine, Scherer, Roesch, & Ellsworth, 2007), and the main conflict is playing out between the negativity and the neutrality position. The negativity position argues from an epistemic perspective that the disconfirmation of expectations is aversive as it signals that one's model of the world does not hold true, which causes cognitive distress (e.g., Carlsmith & Aronson, 1963; Heine, Proulx, & Vohs, 2006; Levy, Harmon-Jones, & Harmon-Jones, 2018; Ludden et al., 2012; Miceli & Castelfranchi, 2002; Noordewier & Breugelmans, 2013). This view is also broadly bolstered by the domain of *cognitive (in)consistency* according to which the conflict between two co-existing relevant cognitions induces negative affect because it runs contrary to the pursuit of cognitive consonance and homeostatis (Festinger, 1957; for an overview, see Gawronski & Brannon, 2019; Gawronski & Strack, 2012; see also Harmon-Jones, Amodio, & Harmon-Jones, 2009). Recent research has expanded the negativity account by offering a process view, suggesting that a mismatch between expectation and event evokes negatively experienced disfluency which may be replaced by further emotions only in a second step after an in-depth event analysis (see Noordewier, Topolinksi, & Van Dijk, 2016; Topolinski & Strack, 2015; however, see also Noordewier & Van Dijk, 2018).

Opposed to this, the neutrality position claims that surprise is not inherently valenced but essentially affect-free and primarily serves as a cognitive marker of the discrepancy between expectation and reality (Lazarus, 1991; Ortony, Clore, & Collins, 1988; Russell, 1980; see also Kruglanski et al., 2018). If at all, surprise provides an undifferentiated burst of arousal that amplifies all subsequent emotional experiences, making you ecstatic of joy for something unexpectedly good and drown in despair for something unexpectedly bad (Mellers, Fincher, Drummond, & Bigony, 2013; Vanhamme, 2003; see also Wilson, Centerbar, Kermer, & Gilbert, 2005).

So much for past unpredicted happenings. But what about unforeseeable future events, and what about the valence of the unpredictable? Current positions again diverge into two main approaches, the ones declaring that people strive to *avoid* the unpredictable, the other ones maintaining that people actively *seek* the unpredictable.

Advocates of the unpredictable-avoidance position argue that humans hold a deeply ingrained craving for certainty and predictability (e.g., Heider, 1958; Kelly, 1955; Miceli & Castelfranchi, 2002). Being able to forecast the future constitutes a fundamental epistemic human need because only an accurate "preview" of the future allows efficient behavior and action control, holding adaptive value in the long run (Gilbert & Wilson, 2007; Thomaschke & Dreisbach, 2013; see also Berlyne, 1960; Schultz, 1998). Whereas every single piece of information that facilitates prediction is experienced as affectively positive and rewarding (Braem & Trapp, 2019; Ogawa & Watanabe, 2011; Trapp, Shenhav, Bitzer, & Bar, 2015), the unpredictable is, in turn, experienced as aversive (see also Arrow, 1965; Pratt, 1964), stressful (de Berker et al., 2016; Peters, McEwen, & Friston, 2017), and anxiogenic (e.g., Grillon, Baas, Lissek, Smith, & Milstein, 2004; Herry et al., 2007; Sarinopoulos et al., 2010; see also Hirsh, Mar, & Peterson, 2012). This psychological preference for the predictable is also largely supported by cognitive science's framework of *predictive coding* that puts the minimization of

free energy (this is: of the discrepancy between prediction and perception) as the overarching principle of every bit of human action (e.g., Clark, 2013; Feldman & Friston, 2010; Friston & Kiebel, 2009; Rao & Ballard, 1999).

But now let's exemplarily turn to Phil Connors, protagonist of the movie "Groundhog Day". Every morning anew, Phil wakes up and runs through the very same day. Caught in this time loop, every situation becomes maximally predictable – however, instead of leaning back and enjoying this state of maximal control, Phil starts to actively disrupt these routines. Why would he do this if predictability was the ultimate goal of his brain? According to the unpredictable-seeking position, not knowing what happens next can also have a strong attractive component: It removes boredom by breaking the dull and colorless daily grind, offers stimulation and variety, and gives that special kick and thrill that paves way for reorientation and change (Berlyne, 1960; McClelland, Atkinson, Clark, & Lowell, 1953; see also Maddi, 1968; Simandan, 2018). Hence, instead of walking into a dark, stimulation-free room and resting there until we die (for an overview on this *dark room problem*, see also Clark, 2018), we sometimes very deliberately expose ourselves to unpredictable events, for instance, by watching mind twist movies, reading a joke, or listening to jazz music (see Przysinda, Zeng, Mayes, Arkin, & Loui 2017; Suls, 1972; Veatch, 1998).

In summary, the psychological landscape on whether unpredictability triggers positive, negative, or neutral affective responses remains heterogeneous. Yet, psychological outcomes result from neurophysiological processes, and maybe these can tell us more about the valence of the unpredicted and the unpredictable.

1.2.2 A Neurophysiological Perspective

Human cognition, emotion, and behavior are a marvel of complexity. Whenever we consciously perceive, feel, and act, this is preceded by a long chain of neurophysiological mechanisms, and comprehending these mechanisms will likewise improve our understanding

of psychological processes. Affective responses cover by themselves a tremendously broad field, and an exhaustive consideration of all possibly involved neuro-components would be out of scope of the current dissertation. For this reason, I will for now focus selectively on the impact of dopamine, in popular science terms also known as the "happy hormone". Being classified as the main actor of the reward system (see Wise, 1980; for an overview, see also Ashby, Isen, & Turken, 1999), it hence presents the probably most intuitive candidate to discuss in the context of positive affect and indeed attracts the largest research interest in the context of unpredictability (however, see also Matias, Lottem, Dugué, & Mainen, 2017, for an overview on the role of serotonin; see Dayan & Yu, 2006; Lauffs, Geoghan, Favrod, Herzog, & Preuschoff, 2020; Yu & Dayan, 2005, for an overview on the role of noradrenaline). A plethora of previous experiments has reliably shown that unpredicted and unpredictable events trigger a release of dopamine (for a recent review, see Diederen & Fletcher, 2020), thereby suggesting that the unpredicted and the unpredictable should be accompanied by positive affect. But this is not the whole story, and as mostly in science, things are not as simple as they seem at first glance. Let us therefore have a closer look on dopamine.

Firstly, there is no such thing as *the* dopaminergic system, but dopamine neurons are primarily located in two systems: (1) the *nigrostriatal system*, which consists of neurons in the *substantia nigra pars compacta* and projects to the movement-associated *striatum*, and (2) the *mesocorticolimbic system*, which consists of neurons in the *ventral tegmental area* and projects to limbic and cortical areas that are primarily involved in reward and motivation (see Arias-Carrión & Pöppel, 2007; Ashby et al., 1999; Diederen & Fletcher, 2020; Schultz, Dayan, & Montague, 1997). Secondly, there is no such thing as *the* dopamine release, but dopamine can be released in different modes. Phasic dopamine release refers to short, only subseconds-lasting, intensive bursts with high amplitudes that predominantly occur in response to external stimuli, whereas tonic dopamine release refers to the continuous and slow background flow of dopamine

that is required for the neural system's normal functioning (e.g., Grace, 1991; see also Schultz, 2016, for a further temporal classification).

According to recent theories from Schultz (2016), phasic dopamine release is again differentiable into two components. The first component acts as a kind of salience detector and responds to anything that is unpredicted, without further distinguishing whether this unpredicted event is rewarding, aversive, novel, or physically intense (see also Ljungberg, Apicella, & Schultz, 1992). The purpose of such a general prediction error alert becomes obvious against the background of adaptational efficiency: Every unpredicted stimulus is potentially important as it could signal reward, threat, or simply a gain of information that might be used to calibrate future behavior and decision-making, thus enabling the organism to learn (e.g., Diederen & Fletcher, 2020; Schultz, 2016). This conceptualization is aptly in line with predictive coding's assumption of a general prediction error that allows the refinement of future predictions until errors no longer occur (see Chapter 1.2.1). The second, slightly later onsetting component is concerned with the actual "content analysis" of the unpredicted event and marks its reward value (Schultz, 2016). It thereby codes the difference between outcome and expectation (Schultz, 1998). This means that in case the event is better than expected, release continues, resulting in an extended rush of dopamine; in case the event is worse than expected, there is a depression of dopamine release; and in case the event is exactly as expected, nothing happens (see also Arias-Carrión & Pöppel, 2007; Schultz et al., 1997; however, see Fiorillo, 2013, for a discussion of dopaminergic processing differences for appetitive vs. aversive prediction errors).

Similar to the psychological temporal stage model presented in Chapter 1.2.1, these findings imply a neurophysiological stage model with initial versus later dopaminergic responses that gradually transition from undifferentiated salience coding towards a more refined

reward value coding that depends on the reward surplus or shortage of the event compared to expectation.

Dopamine is also involved in the processing of unpredictable events; however, much less is known about these dynamics. First evidence implies that in an anticipatory phase before the event, there is a phasic release of dopamine that codes the expected reward value. This phasic neuron release is sensitive to both reward magnitude (i.e., "How much may I get?"), and reward probability (i.e., "How likely will I get this?"), with higher release for higher magnitudes and higher probabilities (e.g., Fiorillo, Tobler, & Schultz, 2003; Preuschoff, Bossaerts, & Quartz, 2006; Tobler, Fiorillo, & Schultz, 2005; Tobler, O'Doherty, Dolan, & Schultz, 2006). In case of not fully (im)probable events (i.e., probabilities of 0% or 100%), there is a slightly later-onsetting and more sustained, tonic release of lower-amplitude dopamine which seems sensitive to variance (this is: to the risk of (not) obtaining the reward), with higher release for higher variances (see Fiorillo et al., 2003; Tobler et al., 2006). This would indicate some per se inherently rewarding qualities of the unpredictable.

So how to finally assess the claim that the unpredicted and the unpredictable should be positively valenced as they are accompanied by dopamine release? The answer may be both "yes" and "no", which may be reconciled by "it depends". In fact, a vast magnitude of findings bolsters the assumption that the function of dopamine goes far beyond – or is even misdescribed – by affect (see Ashby et al., 1999, for the following arguments; see also Wickelgren, 1997; Wise, 2004, 2008). If dopamine was unconditionally linked to positive affect, it should be released for any rewarding event – however, in fact, it is only released for unpredicted rewards. Instead, dopamine is released for a multitude of events that do, per se, not possess any positive or rewarding qualities at all, such as salient, novel, or even stressful or anxiety-inducing events, which would be counterintuitive if dopamine always and only coded positive affect.

The current range of findings rather implies that dopamine has a predominant alert and information function, detecting and pre-categorizing adaptationally potentially relevant events and then projecting them to brain areas that are involved in higher-level cognitive, affective, and motoric processes (e.g., the *hippocampus* for memory, the *prefrontal cortex* for attention, or the *anterior cingulate cortex* for decision-making and modulating emotional responses; see Ashby et al., 1999; Bromberg-Martin, Matsumoto, & Hikosaka, 2010; Diederen & Fletcher, 2020; Schultz, 1998; Schultz & Dickinson, 2000). Dopamine thereby enables cognition, emotion, and behavior in a more complex and interactive manner than its common reputation as a "happy hormone" suggests.

In conclusion, up to now neither psychological nor neurophysiological findings allow a clear derivation of the valence of unpredictability. In the current dissertation, I intend to shed more light on the affective responses to the unpredicted and the unpredictable from a psychological perspective, aiming at uncovering the overarching mechanisms that may subsequently be refined by neurophysiological research. While Chapter 2 contributes to clarifying the valence of the unpredicted, the work presented Chapter 3 takes an in-depth look at the valence of the unpredictable.

1.3 The Current Research

To obtain a comprehensive psychological understanding of unpredictability, it is vital to study this phenomenon from two perspectives: the (already happened) unpredicted and the (yet to happen) unpredictable.

In Chapter 2, I focus on the key determinants and structure of the behavioral, experiential, affective, and cognitive responses to the unpredicted, also known as surprise. Specifically, I empirically contrast the two prevailing, yet competing causal accounts on surprise to conclude whether surprise is rather about unexpectedness or about the ease of making sense of an event. Beyond that, I investigate whether all unpredicted events trigger the

same uniform response patterns, or whether different degrees of deviance and ease of sense-making evoke continuous grades of response intensities. I thereby take a multi-componential view and investigate the behavioral, affective, experiential, and cognitive surprise responses.

In Chapter 3, I quantify the affective value of the (un)predictable by investigating how much of their potential returns people are willing to forgo to ensure interacting with predictable partners. Transferring the economic concepts of risk-return trade-off and risk premium to the psychological domain, I explore the additional expected value of an interaction that is required to make decision-makers indifferent between an unpredictable and a predictable interaction partner. This allows me to extract the valence of the unpredictable.

Combining basic cognitive and socially applied research within an interdisciplinary spectrum of psychology and financial economics, the present work thus provides an integrative, synergetic approach to the responses to and valence of the unpredicted and the unpredictable. This does not only add substantial evidence to current controversial debates, but likewise opens new and innovative research horizons.

Note that Chapter 2 is based on a published manuscript and Chapter 3 is based on a manuscript submitted for publication. As they treat their own topics and questions, both chapters have their own introduction and discussion. In Chapter 4, I present the overarching implications and limitations of the research presented and end with a conclusion.

Chapter 2 – Shades of Surprise: Assessing Surprise as a Function of Degree of Deviance and Expectation Constraints

Abstract

Merging recent surprise theories renders the prediction that surprise is a function of how strong an event deviates from what was expected and of how easily this event can be integrated into the constraints of an activated expectation. The present research investigates the impact of both these factors on the behavioral, affective, experiential, and cognitive surprise responses. In two experiments (total N = 1,257), participants were instructed that ten stimuli of a certain type would appear on the screen. Crucially, we manipulated the degree of deviance of the last stimulus by showing a stimulus that deviated to either no, a medium, or a high degree from the previous nine stimuli. Orthogonally to this deviation, we induced an expectation with either high, moderate, or low constraints prior to the experimental task. We measured behavioral response delay and explicit ratings of liking, surprise, and expectancy. Our findings point out an overall only low association between the behavioral, affective, experiential, and cognitive surprise responses and reveal rather dichotomous response patterns that differentiate between deviance and non-deviance of an event. Challenging previous accounts, the present evidence further implies that surprise is not about the ease of integrating an event with the constraints of an explicit a-priori expectation but rather reflects the automatic outcome of implicit discrepancy detection, resulting from a continuous cognitive fine-tuning of expectations.

2.1 Introduction

Imagine tossing a coin and it lands on edge, finding $100 when tidying up your desk drawers after a long time, or suddenly bumping into your tax consultant at your vacation hotel's

buffet line – we all know the experience that sometimes, there is this intriguing gap between what we think will happen and what actually happens, our ability to predict the course of events breaks down, and we get that distinct feeling of surprise.

Surprise has proven to be more than just short-lived affect. It triggers a whole *emotion syndrome* (Reisenzein, 2000a) of several associated behavioral and mental components, all aiming at facilitating the cognitive mastering of the event and at enabling the organism to respond adaptively to sudden environmental changes (Izard, 1971; Meyer et al., 1991, 1997; Plutchik, 1980). The conscious appraisal of unexpectedness is accompanied by a qualitative feeling of surprise (e.g., Meyer et al., 1991, 1997; Reisenzein, 2000b), ongoing mental and motor activities being interrupted (e.g., Horstmann, 2005; Meyer et al., 1997; Reisenzein, 2000a; Scherer et al., 2004), and we allocate our attention to the surprise-eliciting stimulus (for a recent review and discussion, see Horstmann, 2015). Physiological reactions to surprise comprise an increased skin conductance and a pronounced cardiac response (Niepel, 2001), and even our immediate cortical responses show characteristic reaction patterns to unexpected events (e.g., Donchin, 1981; Ferrari, Bradley, Codispoti, & Lang, 2010; Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003; however, see also Hajcak, Holroyd, Moser, & Simons, 2005). We might express our surprise by spontaneous exclamations (Reisenzein, Meyer, & Niepel, 2012), and according to the culturally associated stereotype, surprise is written in our faces by raised eyebrows, widened eyes, and our mouths slightly opened (e.g., Darwin, 1872; Ekman, 1972, 1979; Ekman & Friesen, 1978; Izard, 1971; however, for a recent discussion, see Reisenzein et al., 2013).

The growing body of evidence on these manifold outcomes of surprise went hand in hand with a revival of the debate on what surprise is essentially. While common knowledge and a long-standing scientific tradition have it that surprise describes the reaction to unexpectedness (e.g., Reisenzein, 2000b; Smedslund, 1990), current research perspectives provide a whole

bunch of models on the causal and procedural architecture of surprise (for a recent overview and discussion, see Reisenzein et al., 2019). The arena of theoretical accounts might be most efficiently characterized by distinguishing two main approaches, the one focusing on expectedness, the other one on explicability and comprehension.

From the classical and lay intuitive perspective, surprise is the response to unexpected events that are discrepant with the schema of a current situation (e.g., Meyer et al., 1991, 1997). A schema is some kind of "mental drawer", representing an organized knowledge structure that guides the processing and integration of incoming information to enable adaptive action control (e.g., Rumelhart & Norman, 1976; Rumelhart & Ortony, 1977). Schemata contain sets of expectations that are used to derive predictions about future events. These expectations can structurally vary in the strength of their constraints, this is, in the typical value range that an event can take (Rumelhart & Norman, 1976). Furthermore, these expectations do not necessarily have to be consciously calculated a priori but can also be implicit or construed after the surprising event. If, for example, a rock flies through your window while you're eagerly writing your manuscript, you will probably not have explicitly expected that *no* rock will fly through your window (unless it is war, or really heavy storm), but you will still be surprised, because your post-hoc assessment of the situation marks the rock as unexpected (Ortony & Patridge, 1987; see also Kahneman & Tversky, 1982; Lorini & Castelfranchi, 2007). Such a perception of "unexpectedness" thus derives from the discrepancy between your expectation and the actual event, giving rise to the initiation of a system-wide surprise mechanism.

The *contrast hypothesis* (Teigen & Keren, 2003; see also Kahneman & Miller, 1986) also builds on the unexpectedness assumption but additionally takes into account the relative strength of disconfirmation and the extent to which the actual event deviates from the expected alternative. According to this hypothesis, surprise is not purely about something being unexpected, but rather about something else being more expected, and the degree of surprise is

derived from the contrast between the prediction and the event. This also resolves the common lay-intuitive conflation of surprise and low probabilities (e.g., Fischhoff, Slovic, & Lichtenstein, 1977; Mellers, Schwartz, & Ritov, 1999): Although most surprising events are also low-probable ones, not every improbable event necessarily elicits surprise because at times, there is no dominant alternative expectation that could be contrasted and disconfirmed. As Teigen and Keren (2003) plausibly illustrate, drawing lottery ticket No. 14,237 from a pool of 100,000 tickets might – despite its low probability – trigger only low surprise unless you strongly expected for example ticket No. 15,031 to be drawn.

Opposed to those accounts, the so-called *sense-making approaches* (Foster & Keane, 2015; see also Maguire & Keane, 2006; Maguire et al., 2011; for a recent discussion, see Reisenzein et al., 2019) do not emphasize the role of expectations but rather conceptualize surprise as resulting from the success or failure to explain an event, with the level of surprise depending on the mental difficulty of integrating it with an existing representation. Coming back to the manuscript-writing scenario, remembering this morning's radio news on a rowdy protest march through your street would probably provide some useful explanation for the rock flying through your window and thus most likely resolve your surprise. This perspective is also supported by findings on the *meaning maintenance model* (e.g., Heine et al., 2006; Proulx & Heine, 2008; Proulx, Heine, & Vohs, 2010; Proulx, Inzlicht, & Harmon-Jones, 2012) implying that expectancy violations trigger the motivation to restore a sense of meaning.

As different as these accounts appear, they share the notion that surprise is (among other things) triggered by an event that deviates from what was expected – and this means regardless of whether it was expected a-priori, post-hoc, implicitly, explicitly, actively or passively (see Lorini & Castelfranchi, 2007; Macedo & Cardoso, 2019; Ortony & Patridge, 1987), or not expected at all (for a similar discussion, see Reisenzein et al., 2012) – or from what was schematically available in a given situation. Logically, this deviation should be continuous, that

is, events might vary in the degree to which they deviate from expectation or situation-related schema activations. In the present paper, we will call this variation in deviance the *degree of deviance* of a surprising event. Although most theories and probably also laypersons share the notion that the degree of deviance, that is, the "surprising-ness" of an event, influences the surprise response (see also Mandler, 1975; Wilson & Gilbert, 2008; that with increasing integration difficulty, affective reactions become more intense), empirical evidence that tests this assertion for directly experienced surprising events is scarce. First valuable evidence was obtained by Foster and Keane (2015) who manipulated the surprising-ness of an event and found that surprise is indeed "a graded experience; it is not all-or-nothing" (p. 75). However, these studies built on surprising events that were encountered by other actors instead of by the participants themselves. The only approach we are aware of that systematically manipulated the degree of deviance of a directly experienced surprising event stems from Reisenzein, et al. (2006; Experiment 1). Thus, the first aim of the present paper was to realize events of varying degree of deviance within a given situation and to gauge whether the degree of deviance affects multiple indicators of surprise.

The second focus of the present paper was *expectation constraints*, that is, the degrees of freedom possible events might have (Rumelhart & Norman, 1976; Rumelhart & Ortony, 1977; see also Schützwohl, 1998). For example, imagine that you are explicitly told – and thus expect – that you will see words on the screen, but actually you see a picture, then you will probably be surprised, because your very specific expectation ("I will see words") was disconfirmed. However, expectations do not necessarily have to be explicit, and surprise can also arise in the absence of any explicitly induced specifications. For instance, being told that you will only see "stimuli" appearing on the screen will not induce an explicit expectation of what specific kind of stimuli will be presented, therefore forming an expectation with lower constraints ("I will see something"). Then, later in the actual task, you will see words, one word after the other, until suddenly, the last stimulus is not a word but a picture. You will also be

surprised, because despite the absence of any specification in the instruction of what the stimuli will be, you incidentally formed an expectation of these "stimuli" obviously being pictures (see Lorini & Castelfranchi, 2007; Ortony & Patridge, 1987; see also Schützwohl, 1998, for further evidence on shaping schema strength by the frequency and variability of schema activations). The question is which way of forming expectations makes you more surprised: being explicitly instructed what type of stimuli to expect (high constraints), or only experiencing these stimuli over the course of the task in absence of an initial explicit expectation (lower constraints)? Building on sense-making approaches according to which more generalized expectations facilitate integration processes and reduce surprise (see Maguire & Keane, 2006; Maguire et al., 2011), we would argue for more surprise with increasing expectation constraints – yet, so far, no one put this case to test.

Now let's go even further: What will happen if an expectation includes a surprise and thus enables the anticipation of a deviant event? Imagine you are told that you will see nine words and a surprising stimulus on the screen – how many instantiations of such a surprise just spring to your mind? Probably a lot: you might see a number, a picture, some abstract shapes, or maybe you might also see an empty screen with no content at all? As these examples point out, by nature, the number of potential events that deviate from a given stimulus profile is larger than the number of events that exactly match the profile. Thus, "expecting the unexpected" should broaden expectations and impose even less constraints: anything that deviates is plausible, and only the non-deviant would be unexpected. From a sense-making perspective, this implies that knowing that something surprising will happen should facilitate integration processes and the cognitive mastering of a surprising event, thereby reducing surprise responses. However, the current research on this issue is inconclusive: Whereas Niepel (2001) reports that announcing a surprising change indeed decreases behavioral surprise responses, Retell, Becker, and Remington (2016) find that explicit information on the appearance of a surprising stimulus does *not* prevent response time interference in a visual search paradigm.

Hence, while previous valuable research has discretely realized different ways of evoking surprise, the present paper is the first to systematically investigate the joint impact of the degree of deviance and expectation constraints on the strength of four different indicators of surprise.

2.2 Aim and design of the present research

The present experiments manipulate the degree of deviance of a surprising stimulus from the stimulus that was expected due to the activation of a specific expectation in a given situation (see also the *contrast hypothesis*, Teigen & Keren, 2003). To illustrate, when you expect to see ten pictures, but the tenth stimulus is either a Chinese ideograph or a non-word, then the non-word deviates from the expectation of a picture very much (a series of letters is definitely not a picture), while the ideograph deviates less so, since the ideograph, although not being a prototypical instantiation of a picture, can be seen as an exotic case of a picture (see also Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Consequently, participants should be more surprised seeing a non-word than an ideograph. Secondly, and orthogonally to that, we manipulated the constraints of a given expectation by either instructing participants explicitly that a certain well-defined category of stimuli would occur (high expectation constraints), or letting them build up an expectation over the course of seeing several stimuli of the same category (moderate expectation constraints), or even telling them that a surprising stimulus would occur (low expectation constraints).

Given previous extensive theorizing as well as layperson conceptions, we predicted that surprise would increase with an increasing degree of deviance of a stimulus and would be higher the more constrained the activated expectation is. Since in the case of explicitly expecting a surprise, only the non-deviant should be surprising, we expected the reversed pattern (decreasing surprise with an increasing degrees of deviance) for lowly-constrained expectations.

As dependent measures, we assessed four indicators of surprise. These were (1) response delay as an indicator of behavioral interruption (e.g., Horstmann, 2005, 2006; Meyer et al., 1991), (2) subjective liking ratings for the surprising stimulus as an indicator of immediate affective responses (e.g., Levy et al., 2018; Noordewier & Breugelmans, 2013; Noordewier et al., 2016; Noordewier & Van Dijk, 2018; Proulx & Inzlicht, 2012; Topolinski & Strack, 2015, for evidence that surprise triggers initial negative affect; see Neta, Norris, & Whalen, 2009, for the affective perception of surprise), (3) subjective surprise ratings as an experiential indicator of surprise (e.g., Reisenzein et al., 2006; Schützwohl, 1998), and (4) subjective expectancy ratings as an indicator of the cognitive appraisal of unexpectedness (e.g., Stiensmeier-Pelster, Martini, & Reisenzein, 1995).

While earlier studies have already implemented surprising events that varied in their strength of surprising-ness (for instance upside-down versus upside-town and Thatcherized faces, see Proulx, Sleegers, & Tritt, 2017; or a slight versus a pronounced change of background patterns, see Reisenzein et al., 2006), and used different approaches to induce experimental expectations (for explicit expectations inductions, see Vanhamme, 2003; for incidental expectation formations, see Horstmann & Schützwohl, 1998; Meyer et al., 1991, 1997; Reisenzein & Studtmann, 2007; for explicit surprise announcements, see Niepel, 2011; Retell et al., 2016), we are not aware of any previous research that orthogonally combined manipulations of the degree of deviance and expectation constraints and measured the (1) behavioral, (2) affective, (3) experiential, and (4) cognitive indicators of surprise within one experiment. Addressing this current research gap allows us to identify the key driving principles for the observed response patterns, thus giving decisive input into the current debate on the determinants of surprise.

Our approach can also shed more light on the long-standing debate to what degree surprise can be conceptualized only in categorical terms of expectancy violation (see also Foster

& Keane, 2015; Maguire et al., 2011; Lorini & Castelfranchi, 2007; Ortony & Patridge, 1987). If surprise was a mere function of unexpectedness, varying degrees of deviance would have no impact on surprise (higher surprise with higher deviance), since both medium- and highly-deviant stimuli are not expected and would thus trigger surprise to a similar degree. The present research puts this absolutistic assumption to test by investigating whether surprise is an all-or-nothing variable, or whether varying degrees of deviance differently affect the strength of the surprise responses.

Data treatment and a priori power-analysis. Due to the novelty of our research question, we could only estimate the effect size and assumed a medium effect of $d = .50$ for the crucial comparison between each two levels of the degree of deviance or expectation constraints. To obtain such an effect in a two-group independent samples *t*-test with an attempted power of .80 and $\alpha = .05$, according to G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), the required sample size is $N = 576$ ($n = 64$ per condition). Resulting from high participant flow, the actual sample sizes slightly exceed these; thus, the present experiments are properly powered.

Due to logistic reasons, data for the three different expectation constraints conditions in Experiment 1 were collected separately and in sequence and thus do not meet the experimental prerequisite of random assignment. However, given the methodological similarity of these experiments (all independent and dependent variables except for the manipulations of expectation constraints were similar), these single sub-experiments are treated as between-subjects conditions in the following for the sake of simplicity. Statistical analyses were computed with *SPSS* Version 25.0 (IBM Corp., 2017). All exclusion criteria of data, all manipulations, all measures, and all preparatory steps prior to the analyses are reported in the text.

2.3 Experiment 1

We developed a new paradigm which allowed us to elicit different grades of surprise by manipulating both the degree of deviance of an event and the constraints of the activated expectation. Building on the *contrast hypothesis* (Teigen & Keren, 2003), we manipulated the degree of deviance by varying the contrast between the expected and the actual event and implemented three stimulus instantiations that conceptually deviated from a current expectation to no degree, to a medium degree, or to a high degree. The challenge in designing such a paradigm was to create a stimulus ecology that contains two strongly contrasting, dissimilar stimulus types, and one intermediate category that overlaps with both the other types. We built on the classic verbal-pictorial distinction that puts verbal and pictorial stimuli as two qualitatively different categorical cues (e.g., Tversky, 1969; Underwood, 1952), and therefore chose non-words and pictures as extreme categories and Chinese ideographs (comprising both verbal and pictorial elements) as the hybrid in between. Thus, seeing a non-word after a series of non-words, or a picture after a series of pictures, should not be surprising at all. However, seeing a picture after a series of non-words, or a non-word after a series of pictures, should be more surprising than seeing a Chinese ideograph after a series of non-words or pictures, since the conceptual deviation is larger.

Orthogonally to that, we systematically manipulated the constraints of the current expectation and induced an expectation which had either high, moderate, or low constraints. This was achieved by implementing different instructions prior to stimulus presentations. To induce an expectation with high constraints, participants read a task instruction that explicitly

announced that pictures (or non-words[2]) would be shown on the screen during the experiment, thus giving very tight restrictions of what to expect. To induce an expectation with moderate constraints, we did not inform participants about the type of the appearing stimuli but only announced unspecific "stimuli", thereby reducing the constraints. As participants were presented a sequence of nine stimuli of the same type until an unannounced stimulus change took place, we argue that this inferior familiarization is sufficient to incidentally form an expectation that then can be violated by a more or less deviating last stimulus (see Ortony & Patridge, 1987; Reisenzein et al., 2019). Lastly, to induce an expectation with low constraints, we realized an instruction that allowed for active surprise anticipation by telling participants that they would see nine pictures (or non-words) and one surprising stimulus on the screen (for a similar manipulation, see Niepel, 2001; Retell et al., 2016).

Having read one of these particular instructions, participants were presented nine expectation-conforming stimuli. Crucially, the tenth stimulus either confirmed the expectation as well, or contrasted it to a medium or high degree. Participants were asked to indicate for each stimulus how much they like it, and, for the last stimulus, to rate how surprising it was and how much they had expected it. We measured behavioral interruption via response delay and the affective surprise response by assessing the perceived valence of the surprise stimulus via spontaneous ratings of liking. The experiential feeling of surprise was measured by subjective surprise ratings, and the cognitive appraisal of unexpectedness by ratings of stimulus expectancy.

[2] In order to enhance the instruction comprehensibility for lay participants, we used in the German instruction materials the German word "Phantasiewörter" (literally, fantasy words). This can be translated into English as "words" and as "non-words" (depending on whether emphasizing the "wordness" nature of the stimuli or the artificiality of them being not real meaningful German words). Since both "phantasy words" and "non-words" denote the same concept of letter strings that look like conventional words but in fact do not have meaning, we decided to label these as "non-words" throughout the manuscript for the sake of terminological accuracy and consistency.

2.3.1 Method

2.3.1.1 Participants. *N* = 711 participants were approached at different parts of the campus of the University of Cologne, Germany, and asked to take part in a larger laptop-based experimental session of unrelated tasks. Informed consent was obtained from all individual participants. Due to a laptop crash, data from two participants could not be saved. Participants who did not terminate the experiment or took part more than once were excluded (high expectation constraints: *n* = 8; moderate expectation constraints: *n* = 19; low expectation constraints: *n* = 3). The final overall sample size was *N* = 679 (M_{age} = 23, SD_{age} = 5; 470 female, 200 male, four gender-diverse; demographic information from five participants was missing, nine participants did not provide information on age). Within each of the three expectation constraints sub-experiments, participants were randomly assigned to one of two baseline stimulus type conditions and one of the three degree-of-deviance conditions (see Appendix B, Table 7, for an overview of the sample characteristics and assignments for each sub-experiment).

2.3.1.2 Materials. We implemented three different stimulus types in our paradigm. These were non-words, pictures, and Chinese ideographs. As non-word stimuli, we employed ten six-letter non-words taken from a pool from Topolinski, Maschmann, Pecher, and Winkielman (2014; see Appendix C for a list of the stimuli employed). Since participants had never encountered these non-words before, and since these stimuli controlled for any systematic influences of language characteristics on affective preferences (see Topolinski et al., 2014), it was assured that liking ratings would not reflect an artifact of confounded linguistic characteristics. Pictorial stimuli were selected from the International Picture Naming Project (IPNP) online database (Szekely et al., 2004) which contains 520 black-and-white drawings of common objects and 275 action presentations. We selected ten stimuli of neutral valence from the syntactic category of *object pictures* and its semantic subcategories *small artifacts*, *large*

artifacts, *vehicles*, *animals*, and *objects and phenomena in nature*. The items we chose displayed a bag, a bench, a bird, a box, a car, a clock, a glass, a lamp, a leaf, and a mirror (see Appendix C). Ten Chinese ideographs were taken from an item pool from Topolinski and Strack (2009; see Appendix C).

2.3.1.3 Apparatus and procedure. Instructions and stimuli were presented against a white background on a 13.3-inch display with a resolution of 1,366 x 768 pixels and a refresh rate of 60 Hz via *DirectRT* software Version 2014.1.123 (Jarvis, 2014). Non-words were presented in Arial font with a size of 6.5 cm X 0.5 cm, and Chinese ideographs and pictures with a size of 4.7 cm X 4.7 cm. Participants were told that the study was a short evaluation task. In the high expectation constraints condition, they were instructed that in the following, they would see ten pictures (or non-words) presented for one second on the screen. In the moderate expectation constraints condition, they were told that ten stimuli would appear on the screen for one second. In the low expectation constraints condition, participants were informed that they would see nine pictures (or non-words) and one more or less surprising stimulus presented for one second on the screen. We counterbalanced between participants which stimulus type (non-words vs. pictures) was announced as the expectable default and presented during the first nine baseline trials. Crucially, we manipulated the degree of deviance of the tenth stimulus between participants and implemented either no, a medium, or a high degree of deviance of this last stimulus (see Figure 1). In the no-deviance condition, the last stimulus was a non-word in the condition in which non-words had been instructed or presented during the baseline trials, and a picture in the condition in which pictures had been instructed or presented during the baseline trials. In the medium-deviance condition, the last stimulus was a Chinese ideograph for both conditions. In the high-deviance condition, the last stimulus was a picture in the condition in which non-words had been instructed or presented during the baseline trials, and a non-word in the condition in which pictures had been instructed or presented during the baseline trials.

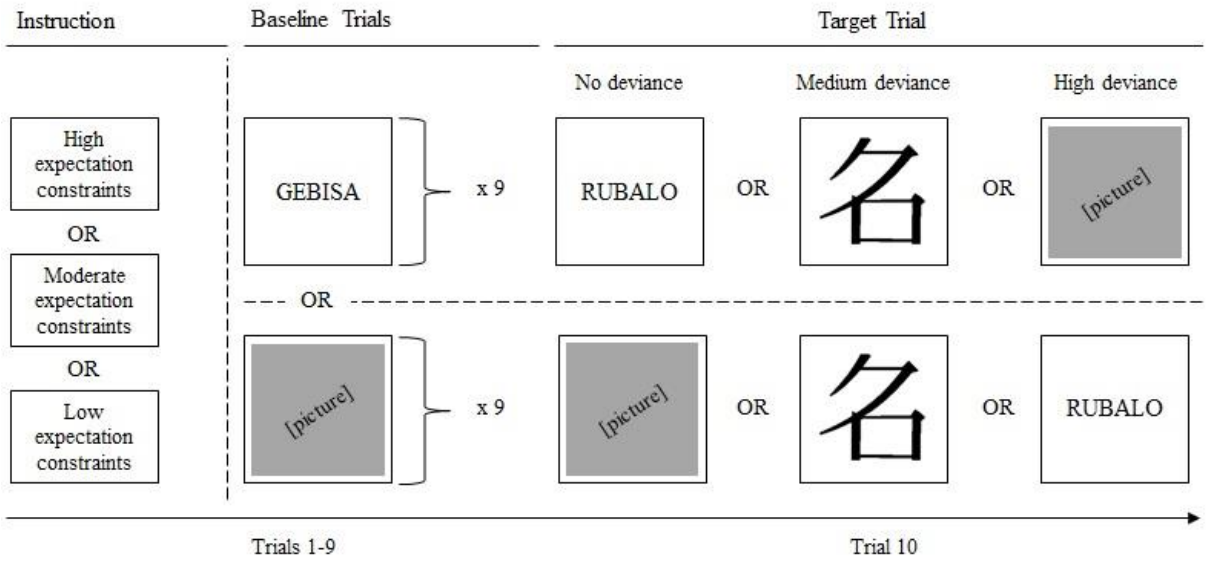

Figure 1. Schematic display of the experimental procedures in Experiment 1. Expectation constraints (high vs. moderate vs. low expectation constraints) were manipulated between three separate sub-experiments. Baseline stimulus type (non-words vs. pictures) and degree of deviance of the tenth target stimulus (no vs. medium vs. high deviance) were counterbalanced between participants in each sub-experiment. Stimuli were randomly drawn from the respective stimulus lists.

The first nine baseline trials were randomized for each participant by drawing a random stimulus from the respective stimulus type list without stimulus replacement. The presentation of the (more or less) deviant stimulus was fixed at the tenth target trial. Stimuli were presented in the center of the screen for 1,000 ms. Participants were asked to indicate for each stimulus how much they like it on a scale from zero (*not at all*) to ten (*very much*). Upon key press and after an inter-trial interval of 1,000 ms, the next trial followed. For the crucial last stimulus, participants were additionally asked to judge on a scale from zero (*not at all*) to ten (*very much*) how much they were surprised by this item, and how much they had expected it. The assessment order of ratings of surprise and expectancy was fixed. Thus, participants completed the following rating sequence: (1) liking ratings for the nine baseline stimuli, (2) liking rating for

the tenth stimulus, (3) surprise rating for the tenth stimulus, (4) expectancy rating for the tenth stimulus.

2.3.2 Results

Trials with invalid responses (i.e., non-numeric or exceeding the scale; 17 of 8,148; 0.21 %) and liking rating trials with reaction times exceeding 10,000 ms (230 of 6,790; 3.39 %) were excluded to prevent an influence of extreme outliers and to ensure the assessment of spontaneous affective responses. Since we were a priori interested in how expectation constraints and the degree of deviance of the last stimulus would affect each of the four surprise syndrome components, we performed a univariate 3 (Expectation constraints: high vs. moderate vs. low expectation constraints; between-participants) X 3 (Degree of deviance of the last stimulus: no vs. medium vs. high deviance; between-participants) X 2 (Baseline stimulus type: non-words vs. pictures; between-participants) ANOVA separately for each syndrome component measure and refrained from conducting omnibus analyses for the ease of interpretability. This presents a common procured in the existing surprise literature (e.g., Meyer et al., 1997; Niepel, 2001; Schützwohl, 1998). Baseline stimulus type effects are reported only when significant. Significant effects were investigated with Bonferroni-corrected independent-samples *t*-tests (Bonferroni-adjusted $p \leq .025$, or $p \leq .017$, depending on the number of tests required for testing the specific hypothesis), reporting adjusted degrees of freedom in case of variance heterogeneity as indicated by statistical significance of Levene's Test for Equality of Variances. The reported confidence intervals (95% $CI_{difference}$) indicate the confidence interval for the difference between the means of the groups that are compared and not to the confidence interval for the effect. Table 1 provides an overview on the means and standard deviations for all measures and conditions. Furthermore, the joint assessment of these four surprise components allowed correlational analyses (Bonferroni-adjusted $p \leq .008$).

Table 1

Experiment 1: Means and standard deviations for the four assessed surprise syndrome components for no, medium, and high degree of stimulus deviance for high, moderate, and low expectation constraints. Results for the non-word baseline condition are reported in the upper row. Results for the picture baseline condition are reported in the lower row.

		Expectation constraints					
		High		Moderate		Low	
Degree of deviance	Syndrome component	*M*	*SD*	*M*	*SD*	*M*	*SD*
No deviance	Response delay[a]	-418 7	1295 1106	-158 138	1384 1446	-13 344	1644 1611
	Liking ratings[b]	-0.03 -0.10	2.51 2.71	0.00 -0.14	2.65 3.50	-0.20 -0.13	2.62 3.09
	Surprise ratings	3.78 4.29	2.64 3.37	2.55 2.70	2.68 2.54	3.50 2.70	2.68 2.74
	Expectancy ratings	4.00 3.95	2.61 3.40	3.76 3.02	2.83 2.87	4.25 4.08	3.12 2.88
Medium deviance	Response delay[a]	2036 641	1806 1554	1353 539	1732 2016	1635 457	2125 1410
	Liking ratings[b]	0.19 -0.68	3.40 3.52	0.54 -0.43	2.95 3.91	0.56 -2.71	2.76 3.52
	Surprise ratings	7.45 6.12	2.26 3.10	7.83 7.89	2.44 1.90	7.13 6.34	2.42 3.24
	Expectancy ratings	2.40 2.29	3.34 2.60	2.28 2.57	2.89 2.85	3.51 2.92	3.40 3.39
High deviance	Response delay[a]	2939 2042	1984 1925	1924 2088	1883 1523	2441 2030	2122 2065
	Liking ratings[b]	0.11 -1.63	3.68 3.67	1.37 -1.41	2.75 3.15	0.42 -2.38	3.58 3.20
	Surprise ratings	8.03 7.34	2.39 2.68	8.17 6.78	2.33 3.11	6.10 6.16	3.48 3.38
	Expectancy ratings	2.03 1.97	2.76 3.07	2.14 3.28	2.83 3.08	2.46 1.65	2.99 2.62

[a]Response delay is reported as the difference between the reaction time for the liking rating for the tenth target stimulus and the reaction time for the liking rating for the ninth baseline stimulus (in milliseconds).

[b]Liking ratings are reported as the difference between the liking ratings for the tenth target stimulus and the ninth baseline stimulus.

2.3.2.1 Response delay. Since liking ratings were the first behavioral response assessed, reaction times for liking ratings were used as an index of response delay. To establish an effect measure that controls for intra-individual variance and habituation effects throughout the experimental procedure, we calculated the difference between the liking reaction time for the tenth target stimulus and the liking reaction time for the immediately prior ninth baseline stimulus for each participant. This measure thus indicates how much longer participants needed to respond to the final expectation-(non)-conforming stimulus compared to the expectation-conforming preceding stimulus.

The ANOVA revealed a significant main effect of the degree of deviance, $F(2,644) = 95.78$, $p < .001$, $\eta_p^2 = .23$, but no significant main effect of expectation constraints, $F(2,644) = 1.02$, $p = .36$, and no significant interaction between the two factors, $F(4,644) = 1.18$, $p = .32$. We also obtained a significant main effect of the baseline stimulus type, $F(1,644) = 8.19$, $p = .004$, $\eta_p^2 = .23$, and a significant interaction between the degree of deviance and the baseline stimulus type, $F(2,644) = 10.43$, $p < .001$, $\eta_p^2 = .03$. Simple main effects analyses indicated that the main effect of the degree of deviance was significant for both baseline stimulus type conditions (both Fs ≥ 36.80, both ps $> .001$). In the non-word baseline stimulus type condition, the response delay was stronger for medium-deviant stimuli compared to non-deviant stimuli, $t(198.18) = 8.17$, $p < .001$, 95% $CI_{difference} = [1416, 2317]$, $d = 1.10$. The response delay was also stronger for highly- compared to non-deviant stimuli, $t(193.14) = 11.04$, $p < .001$, 95% $CI_{difference} = [2155, 3093]$, $d = 1.49$, and compared to medium-deviant stimuli, $t(213) = 2.83$, $p = .005$, 95% $CI_{difference} = [229, 1285]$, $d = 0.39$. In the picture baseline stimulus type condition, there was no significant difference in the response delay between medium- and non-deviant

stimuli, $t(211.31) = 1.86$, $p = .07$. However, the response delay was stronger for highly-deviant stimuli compared to non-deviant stimuli, $t(195) = 8.60$, $p < .001$, 95% $CI_{difference} = [1457, 2324]$, $d = 1.16$, and compared to medium-deviant stimuli, $t(213) = 6.32$, $p < .001$, 95% $CI_{difference} = [1039, 1982]$, $d = 0.86$.

2.3.2.2 Liking ratings. Similarly, to analyze the relative liking of the target stimulus, we calculated the difference between the liking rating for the tenth target stimulus and the liking rating for the ninth baseline stimulus for each participant. Thus, this measure indicates how much more participants liked the crucial last stimulus compared to the preceding expectation-conforming stimulus.

The ANOVA neither revealed a significant effect of the degree of deviance, $F(2,644) = 1.35$, $p = .26$, nor a significant main effect of expectation constraints, $F(2,644) = 2.87$, $p = .06$, nor a significant interaction between the two, $F(4,644) = 0.73$, $p = .57$. The main effect of the baseline stimulus type was significant, $F(1,644) = 31.43$, $p < .001$, $\eta_p^2 = .05$, and we also observed a significant interaction between the degree of deviance and the baseline stimulus type, $F(2,644) = 8.27$, $p < .001$, $\eta_p^2 = .03$. Simple main effects analyses indicated that the effect of the degree of deviance was not significant for the non-word baseline stimulus type condition, $F(2,644) = 1.46$, $p = .23$, whereas it was significant for the picture baseline stimulus type condition, $F(2,644) = 8.21$, $p < .001$, $\eta_p^2 = .03$: When presenting pictures in the first nine trials, medium-deviant stimuli received relatively lower liking ratings than non-deviant stimuli, $t(208.76) = -2.58$, $p = .011$, 95% $CI_{difference} = [-2.09, -0.28]$, $d = -0.34$, and also highly-deviant stimuli were liked relatively less than non-deviant stimuli, $t(222) = -3.92$, $p < .001$, 95% $CI_{difference} = [-2.53, -0.84]$, $d = -0.52$. We did not obtain a significant difference for the liking ratings between highly- and medium-deviant stimuli, $t(213) = -1.03$, $p = .31$.

2.3.2.3 Surprise ratings. We found a significant main effect of the degree of deviance, $F(2,661) = 147.49$, $p < .001$, $\eta_p^2 = .31$, a significant main effect of expectation constraints,

$F(2,661) = 5.90$, $p = .003$, $\eta_p^2 = .02$, and a significant interaction between the two factors, $F(4,661) = 4.99$, $p = .001$, $\eta_p^2 = .03$. The ANOVA also revealed a significant main effect of the baseline stimulus type, $F(1,661) = 4.83$, $p = .028$, $\eta_p^2 = .01$, which did, however, no longer reach significance in an independent-samples *t*-test for the surprise ratings between the non-word versus picture baseline stimulus type condition, $t(677) = 1.86$, $p = .06$, and which is conceptually irrelevant. Simple main effects analyses indicated that the main effect of the degree of deviance emerged across all three expectation constraints conditions (all Fs ≥ 36.23, all ps $< .001$). Surprise ratings were significantly higher for medium-deviant stimuli than for non-deviant stimuli (all ts ≥ 5.85, all ps $< .001$), and they were also significantly higher for highly- than for non-deviant stimuli (all ts ≥ 6.14, all ps $< .001$). Surprise ratings did not significantly differ between highly- and medium-deviant stimuli in any of the expectation constraints conditions after Bonferroni-correcting for multiple comparisons (all ts ≤ 1.98, all ps $\geq .05$).

Simple main effects analyses further indicated that the main effect of expectation constraints was significant for all three degree-of-deviance conditions (all Fs ≥ 3.87, all ps $\leq .021$). Follow-up independent-samples *t*-tests showed mixed patterns. For non-deviant stimuli, surprise ratings were significantly higher for highly-constrained expectations compared to moderately-constrained expectations, $t(152) = 3.12$, $p = .002$, 95% $CI_{difference} = [0.52, 2.33]$, $d = 0.50$. The rating difference between highly- and lowly-constrained expectations and between moderately- and lowly-constrained expectations was not significant (both ts ≤ 2.08, both ps $\geq .04$). For medium-deviant stimuli, surprise ratings were significantly lower for highly- than for moderately-constrained expectations, $t(134.71) = -2.53$, $p = .012$, 95% $CI_{difference} = [-1.86, -0.29]$, $d = -0.42$. Whereas there was no significant difference in surprise ratings between highly- and lowly-constrained expectations, $t(147) = 0.17$, $p = .86$, surprise ratings were significantly higher for moderately- compared to lowly-constrained expectations, $t(141.23) = 2.72$, $p = .007$, 95% $CI_{difference} = [0.31, 1.94]$, $d = 0.44$. For highly-deviant stimuli, surprise ratings were not significantly different between highly- and moderately-constrained expectations, $t(145) = 0.50$,

p = .62, but they were significantly higher for highly- than for lowly-constrained expectations, $t(139.17) = 3.19$, p = .002, 95% $CI_{difference}$ = [0.59, 2.52], d = 0.52, and for moderately- compared to lowly-constrained expectations, $t(143.14) = 2.60$, p = .010, 95% $CI_{difference}$ = [0.32, 2.35], d = 0.43.

2.3.2.4 Expectancy ratings. The ANOVA revealed a significant main effect of the degree of deviance, $F(2,660) = 17.46$, $p < .001$, $\eta_p^2 = .05$, but no significant main effect of expectation constraints, $F(2,660) = 1.00$, p = .37, and no significant interaction between the two factors, $F(4,660) = 1.79$, p = .13. Across all expectation constraints conditions, medium-deviant stimuli received lower expectancy ratings than non-deviant stimuli, $t(453) = -4.15$, $p < .001$, 95% $CI_{difference}$ = [-1.74, -0.62], d = -0.39, and also highly-deviant stimuli received lower expectancy ratings compared to non-deviant stimuli, $t(455) = -5.83$, $p < .001$, 95% $CI_{difference}$ = [-2.14, -1.06], d = -0.55. There was no significant rating difference between medium- and highly-deviant stimuli, $t(442) = 1.48$, p = .14.

2.3.2.5 Correlational analyses. Based on the notion that surprise triggers response delay and negative affect, induces an experiential feeling of surprise, and evokes an appraisal of unexpectedness, we assumed positive correlations between response delay and surprise ratings and between liking ratings and expectancy ratings. Conversely, correlations between response delay and liking ratings, response delay and expectancy ratings, liking ratings and surprise ratings, and surprise ratings and expectancy ratings were supposed to be negative. To assess the overall strength of intercorrelation of the four measures we calculated bivariate Pearson correlation coefficients on the aggregated subject-level data. Since we were interested in the strength of association, we focused on the size of the correlation coefficients rather than on their significances (see also Reisenzein, 2000a), but we will report significance patterns (Bonferroni-adjusted $p \leq .008$) for the sake of completeness. As can be seen in Table 2, most correlations were in the predicted direction, but the overall strength of association between the different

syndrome components was only moderate to low. Significant correlations were obtained between response delay and surprise ratings and between expectancy ratings and surprise ratings. Strikingly, correlations involving liking ratings were almost consistent with zero.

Table 2

Experiment 1: Pooled Pearson correlations among the four assessed surprise syndrome components. Statistically significant results are marked with an asterisk ($p \leq .008$, Bonferroni-corrected).

Syndrome component	1	2	3	4
1. Response delay	-			
2. Liking ratings	.00	-		
3. Surprise ratings	.30*	-.03	-	
4. Expectancy ratings	-.14*	.00	-.20*	-

2.3.3 Discussion

In this first experiment, we investigated the impact of the degree of deviance of an event and of expectation constraints on the strength of the behavioral, affective, experiential, and cognitive surprise responses. In line with our predictions, the observed response delay pattern suggests a differentiated sensitivity to the degree of deviance on the immediate behavioral level, pointing out a significantly stronger response delay for increasingly deviant events. The non-significant difference in the response delay between non- and medium-deviant events in the picture baseline condition may be attributable to random fluctuations, since we did not observe systematic variations caused by the baseline stimulus type.

Liking ratings were strongly affected by an interaction between the degree of deviance and the baseline stimulus type. Whereas the non-word baseline condition remained unaffected

by the degree of deviance, the picture baseline condition revealed a dichotomy of liking ratings in the predicted direction, with (both medium- and highly-) deviant stimuli being liked relatively less than non-deviant stimuli. We think that this pattern dissociation might be explained by a confound between affective responses triggered by surprising stimuli, and affective responses triggered by stimulus-inherent characteristics. Despite being neutral in valence, the picture stimuli employed in the current experiment held meaningful information (by, e.g., depicting a bird, or a car), thus likely eliciting stronger affective responses than the contentually meaning*less* non-words and ideographs (see also Hinojosa, Carretié, Valcárcel, Méndez-Bértolo, & Pozo, 2009; Kensinger & Schacter, 2006; Larsen, Norris, & Cacioppo, 2003, for a discussion of affective processing differences between verbal and pictorial stimuli). Hence, seeing a non-word or a Chinese ideograph after a series of pictures may not only cause a liking decrease due to surprise, but also due to overall reduced affective preferences for these comparatively rather "boring" stimuli[3]. Assuming an interplay between these two processes can also explain why we did not observe a significant liking increase for pictures following the last non-word baseline trials, since the surprise-induced negativity effect might have counteracted the picture-induced positivity effect.

For the more cognitive measures of surprise ratings and expectancy ratings, our findings point towards a rather categorical, dichotomous clustering of events into "non-surprising/expected" and "surprising/unexpected", with rating differences emerging only when comparing non-deviant to deviant stimuli, but not when comparing deviant stimuli that deviate to varying degrees. It was only for the surprise ratings that we found an impact of expectation constraints, however, the non-systematic patterns did not support our hypothesis of increasing surprise ratings with increasing expectation constraints. Remarkably, we also did not find

[3] Exploratory analyses comparing the liking ratings between pictures and non-words across all nine baseline trials indicate that pictures were indeed liked significantly more than non-words, $t(677) = 6.29$, $p < .001$, 95% $CI_{difference}$ [0.42, 0.80], $d = 0.48$, suggesting a general affective preference for pictorial stimuli. For Chinese ideographs, the calculation of a corresponding contrast was not possible because this stimulus type was never employed during the baseline.

evidence for a reversal of response patterns for expectations with low constraints that explicitly comprised the occurrence of a surprise. Thus, making the unexpected initially highly expected does not seem sufficient to prevail against dynamic expectancy formation during the baseline trials. We will come back to discussing this implication in more detail in the General Discussion. The overall strength of association between the four syndrome components was only weak, which is in line with previous research (Reisenzein, 2000a).

Summarizing the above-described results, it seems that our choice of stimuli turns out as a potential shortcoming. Given the impact of the baseline stimulus type on liking ratings and the theoretical restraints of deploying both verbal and pictorial stimuli within one experiment (see also Clark & Paivio, 1991; De Houwer & Hermans, 1994, for a general account of processing fragmentation for verbal vs. non-verbal stimuli), we thus cannot be certain whether the observed findings indeed reflect effects of the degree of deviance, or whether they are (at least partly) caused by stimulus characteristics. To overcome these flaws and to obtain clearer results, we decided to conduct a second experiment that controlled for the effectiveness of our manipulations and for potential stimulus effects.

2.4 Experiment 2

Experiment 2 deployed two verbal stimulus types (letters and one-digit numbers) that were perceptually comparable and both semantically and affectively neutral. To induce different degrees of deviance, we implemented a second stimulus variable, namely the color of the stimuli appearing on the screen (see Horstmann, 2002, 2005, 2006, for previous surprise manipulations using color changes). In order to control for strong color effects and to ensure that both colors were encodable in a similar way, we decided to employ the (physically and psychologically) closely related colors green and blue. Given this type–color matrix, the overall stimulus expectancy should be composed of two distinct expectancies: one regarding the type, and one regarding the color of the stimuli presented on the screen (for similar considerations,

see Macedo, Cardoso, Reisenzein, Lorini, & Castelfranchi, 2009; Reisenzein et al., 2019). Thus, given the example of initially instructed green letters, a final blue letter would be surprising, but a final blue number would be even more surprising. To ensure that these new stimuli were perceived as varying in deviance, and to test whether the instructions effectively induced expectations with different constraints, we conducted two pretests prior to the main experiment. Furthermore, we implemented a slight change in our measurement of the experiential feeling of surprise by directly asking participants how much they *felt* surprised (instead of how much they *were* surprised, as in Experiment 1) to narrow the scope of the assessment more clearly to the experiential surprise component[4].

2.4.1 Method

2.4.1.2 Participants. N = 586 participants were recruited from different parts of the campus of the University of Cologne, Germany, and asked to take part in a larger computer-based experimental session of multiple unrelated tasks. All participants indicated their informed consent. Data from two participants could not be saved due to technical issues. Participants who did not terminate the experiment were excluded (n = 4). We further excluded data from two participants who started a conversation during the experiment. The final overall sample size was N = 578 (M_{age} = 23, SD_{age} = 4; 323 female, 243 male, 8 gender-diverse; 5 participants did not report full demographics). Participants were randomly assigned to one of the three degree-of-deviance conditions, one of the three expectation-constraints conditions, one of the two baseline stimulus type conditions, and one of the two baseline stimulus color conditions (for an overview, see Appendix B, Table 8).

2.4.1.3 Materials. We implemented two different stimulus types and two different stimulus colors. As stimulus types, we employed the 26 basic uppercase letters from the German alphabet, and the ten one-digit numbers from 0 to 9. As stimulus colors, we chose the colors

[4] We thank an anonymous reviewer for this suggestion.

green (*DirectRT* styles color code 50708) and blue (*DirectRT* styles color code 12615680). In line with previous surprise research employing color manipulations (Horstmann, 2002, 2005, 2006) we did not match the colors for luminance but instead considered participants' subjectively perceived differences as sufficient.

2.4.1.4 Apparatus and procedure. Instructions and stimuli were presented in the center of the screen against a white background on a 13.3-inch display with a resolution of 1,366 x 768 pixels and a refresh rate of 60 Hz via *DirectRT* software Version 2014.1.123 (Jarvis, 2014). Letter and number stimuli were presented in Arial font with a size of about 0.8 cm X 1.0 cm. Participants were told that the study was a short evaluation task. Similar to Experiment 1, in the high expectation constraints condition, participants were told that they would see ten green letters, green numbers, blue letters, or blue numbers (depending on the condition) presented on the screen for one second. In the moderate expectation constraints condition, they were instructed that in the following, ten stimuli would appear for on the screen for one second. In the low expectation constraints conditions, participants were informed that they would see nine green letters, green numbers, blue letters, or blue numbers (depending on the condition) and one surprising stimulus presented for one second on the screen.

We manipulated the degree of deviance by varying the number of simultaneous stimulus expectancy disconfirmations (none vs. one vs. both stimulus expectancies disconfirmed). This was operationalized by the number of stimulus type and color changes between the baseline stimuli and the target stimulus. As in the first experiment, we implemented either no, a medium, or a high degree of deviance of the target stimulus (see Figure 2). In the no-deviance condition, the last stimulus was exactly of the type and color as the preceding baseline stimuli. In the medium-deviance condition, the color of the last stimulus changed from green to blue and vice versa, depending on the color of the baseline stimuli. In the high deviance condition, the last stimulus was of a different type and color as the preceding baseline stimuli. We refrained from

also realizing a same-color-different-type manipulation of medium deviance for reasons of stimulus dominance effects: While the stimulus type was announced as the salient category in the instructions (e.g., “You will see ten green letters”), information on stimulus color was only mentioned as a subordinate attribute describing this stimulus type and thus of minor information (for an overview on the semantically superordinate position of nouns compared to adjectives, see Carnaghi, Maass, Gresta, Bianchi, Cadinu, & Arcuri, 2008; see also Furtner, Rauthmann, & Sachse, 2009, for evidence on the dominance of nouns in German linguistic processing). A change of stimulus type might thus have been more salient than a change in color, leading to a potential skew in the deviance perception from medium to rather high.

The sequence of nine baseline trials was randomized for each participant by drawing a random stimulus from the respective stimulus type list and presenting it in the respective color. Baseline stimulus type and baseline stimulus color were counterbalanced between participants. The presentation of the (more or less) deviant target stimulus was fixed at the tenth trial, with the non-, medium-, or highly-deviant stimulus type being drawn from the respective stimulus type list and presented in the respective color. Stimuli were drawn from the lists without replacement and presented in the center of the screen for 1,000 ms. Participants were asked to indicate for each stimulus how much they like it on a scale from zero (*not at all*) to ten (*very much*). Upon key press and after an inter-trial interval of 1,000 ms, the next trial followed. For the crucial tenth target stimulus, participants were additionally asked to judge on a scale from zero (*not at all*) to ten (*very much*) how much they felt surprised by this stimulus, and how much they had expected it. The trial order of ratings of surprise and expectancy was fixed, following the procedures of Experiment 1.

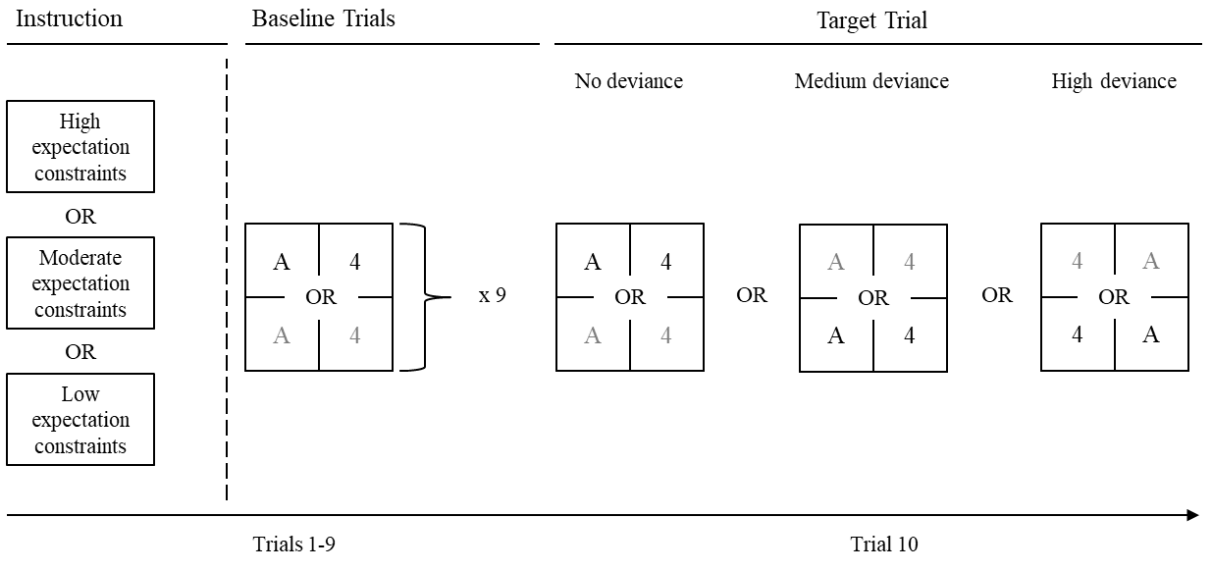

Figure 2. Schematic display of the experimental conditions in Experiment 2. Expectation constraints (high vs. moderate vs. low expectation constraints), baseline stimulus type (green vs. blue, letters vs. numbers; the different colors are depicted via grey and black in the present figure) and degree of deviance of the tenth target stimulus (no vs. medium vs. high deviance) were counterbalanced between participants. Stimuli were randomly drawn from the respective stimulus lists. Each of the fours boxes presents one of the four possible baseline conditions and the resulting deviance scenarios.

2.4.2 Pretests

We conducted two manipulation check pretests prior to the main Experiment 2 to test the effectiveness of our manipulations of the degree of deviance and of expectation constraints.

2.4.2.1 Degree of deviance. N = 67 participants (M_{age} = 23, SD_{age} = 5; 49 female, 16 male, 2 gender-diverse; two participants did not provide information on age) were asked to assess the similarity between each of the stimulus combinations employed for the degree-of-deviance manipulation on a scale from zero (*not similar at all*) to ten (*very similar*). Crucially, we expected that non-deviance stimulus combinations (same type, same color) would be rated

as being significantly more similar to each other than medium-deviance stimulus combinations (same type, different color), and than high-deviance stimulus combinations (different type, different color), and that medium-deviance stimulus combinations would be rated as being significantly more similar to each other than high-deviance stimulus combinations. Table 3 provides an overview on the means and standard errors for the similarity ratings. Paired-samples *t*-tests (Bonferroni-adjusted $p \leq .017$) revealed that all effects were significant in the predicted direction (all ts ≥ 4.45, all ps $\leq .001$).

Table 3

Pretest – Degree of deviance: Means and standard errors for the similarity ratings for no-, medium-, and high-deviance stimulus combinations (no deviance: same stimulus type, same stimulus color; medium deviance: same stimulus type, different stimulus color; high deviance: different stimulus type, different stimulus color).

		Letters		Numbers	
		Blue	Green	Blue	Green
Letters	Blue	*M* = 6.31, *SE* = 0.33	–	–	–
	Green	*M* = 4.40, *SE* = 0.33	*M* = 6.30, *SE* = 0.32	–	–
Numbers	Blue	[a]	*M* = 2.62, *SE* = 0.27	*M* = 6.45, *SE* = 0.33	–
	Green	*M* = 3.08, *SE* = 0.36	[a]	*M* = 4.10, *SE* = 0.33	*M* = 6.45, *SE* = 0.29

[a]Moderate deviance conditions were only realized by changing the stimulus color, not by changing the stimulus type. Hence, similarity ratings for different-type-same-color stimulus combinations were not assessed.

2.4.2.2 Expectation constraints. N = 108 participants (M_{age} = 23, SD_{age} = 4; 32 female, 73 male, 2 gender-diverse; demographic data from one participant was missing; two

participants did not provide information on age) were randomly assigned to one of the two baseline stimulus type conditions and one of the two baseline stimulus color conditions. Degree of deviance and expectation constraints were manipulated within-participants. Each participant was presented a random sequence of nine trials, with each trial combining one specific expectation constraints manipulation with one specific degree-of-deviance manipulation. As stimuli, we employed the stimulus material that was successfully validated in the degree-of-deviance pretest. In each trial, participants were presented one of the specific instructions and were then given hypothetical information on the type and color of the ten stimuli appearing on the screen, with the tenth stimulus deviating from the previous nine stimuli to an either high degree, to a medium degree, or not at all. After reading this information, they were asked to indicate how much one would expect the respective tenth stimulus on a scale from zero (*not at all*) to ten (*very much*). We expected a decrease in expectancy ratings with increasing expectation constraints for both highly- and medium-deviant stimuli. For non-deviant stimuli, we tentatively expected the reverse pattern, with expectancy ratings increasing with increasing expectation constraints. To test these crucial hypotheses, we conducted paired-samples *t*-tests (Bonferroni-adjusted $p \leq .006$) comparing the expectancy ratings between the three expectation constraints conditions separately for each degree-of-deviance condition. The results indicated that all effects were significant in the predicted direction, thus confirming the effectiveness of our manipulation (all ts ≥ 2.97, all ps $\leq .003$). Table 4 depicts the means and standard errors for the expectancy ratings.

Table 4

Pretest – Expectation constraints: Means and standard errors for the expectancy ratings for no, medium, and high degree of stimulus deviance for high, moderate, and low expectation constraints for each baseline condition.

		Expectation constraints		
		High	Moderate	Low
	Baseline cond.			
No deviance	Green numbers	*M* = 6.82, *SE* = 0.76	*M* = 6.29, *SE* = 0.64	*M* = 5.57, *SE* = 0.73
	Blue numbers	*M* = 7.65, *SE* = 0.65	*M* = 6.65, *SE* = 0.70	*M* = 4.46, *SE* = 0.73
	Green letters	*M* = 8.96, *SE* = 0.49	*M* = 7.41, *SE* = 0.52	*M* = 4.07, *SE* = 0.71
	Blue letters	*M* = 7.22, *SE* = 0.77	*M* = 6.59, *SE* = 0.68	*M* = 3.70, *SE* = 0.62
Medium deviance	Green numbers	*M* = 2.71, *SE* = 0.54	*M* = 3.21, *SE* = 0.47	*M* = 7.04, *SE* = 0.56
	Blue numbers	*M* = 3.39, *SE* = 0.66	*M* = 5.23, *SE* = 0.61	*M* = 7.85, *SE* = 0.49
	Green letters	*M* = 2.78, *SE* = 0.61	*M* = 5.15, *SE* = 0.57	*M* = 7.41, *SE* = 0.40
	Blue letters	*M* = 1.96, *SE* = 0.53	*M* = 5.26, *SE* = 0.56	*M* = 7.48, *SE* = 0.50
High deviance	Green numbers	*M* = 2.36, *SE* = 0.58	*M* = 3.61, *SE* = 0.50	*M* = 6.50, *SE* = 0.61
	Blue numbers	*M* = 3.65, *SE* = 0.74	*M* = 5.08, *SE* = 0.68	*M* = 7.04, *SE* = 0.58
	Green letters	*M* = 2.33, *SE* = 0.65	*M* = 4.67, *SE* = 0.63	*M* = 6.96, *SE* = 0.60
	Blue letters	*M* = 2.22, *SE* = 0.62	*M* = 4.04, *SE* = 0.53	*M* = 6.44, *SE* = 0.60

2.4.3 Results

As in Experiment 1, we excluded trials with invalid responses (i.e., non-numeric or exceeding the scale; 11 of 6,936; 0.16 %) and liking rating trials with reaction times exceeding 10,000 ms (178 of 5,780; 6.80 %). Because the high comparability of the stimuli employed in Experiment 2 does not give a priori reasons for assuming substantial stimulus effects, we refrained from including the factors baseline stimulus type and baseline stimulus color into the

analyses in order to maintain high power and to reduce the inflation of false positives[5]. We performed a univariate 3 (Expectation constraints: high vs. moderate vs. low expectation constraints; between-participants) X 3 (Degree of deviance of the last stimulus: no vs. medium vs. high deviance; between-participants) ANOVA separately for each of the four components. Follow-up comparisons and correlational analyses were conducted as in Experiment 1. Table 5 provides an overview on the means and standard deviations for all measures and conditions.

[5] Supplementary analyses including both baseline stimulus type and baseline stimulus color can be found on https://osf.io/68du7/. In a nutshell, we did not find evidence for systematic influences of baseline stimulus type and color. Occasional effects might also be false positives, resulting from the high number of tests performed that were not taken into account in the a priori power analyses.

Table 5

Experiment 2: Means and standard deviations for the four assessed surprise syndrome components for no, medium, and high degree of stimulus deviance for high, moderate, and low expectation constraints.

		Expectation constraints					
		High		Moderate		Low	
Degree of deviance	Syndrome component	*M*	*SD*	*M*	*SD*	*M*	*SD*
No deviance	Response delay[a]	-94	1527	-211	1912	222	1649
	Liking ratings[b]	0.78	3.60	-0.31	2.82	-0.17	3.05
	Surprise ratings	3.30	2.96	2.78	2.38	2.07	2.41
	Expectancy ratings	3.41	2.94	5.17	3.32	4.60	3.51
Medium deviance	Response delay[a]	1406	2004	779	1501	1164	1689
	Liking ratings[b]	-0.54	3.30	0.18	2.78	0.82	3.74
	Surprise ratings	5.11	3.64	4.86	3.31	4.97	3.23
	Expectancy ratings	2.46	3.16	3.21	2.60	3.06	3.11
High deviance	Response delay[a]	2457	2531	1136	1447	1190	1402
	Liking ratings[b]	-0.48	3.30	0.29	2.77	-0.69	2.98
	Surprise ratings	6.66	3.09	5.35	3.33	4.60	3.06
	Expectancy ratings	2.13	3.14	2.55	3.13	3.44	3.29

[a]Response delay is reported as the difference between the reaction time for the liking rating for the tenth target stimulus and the reaction time for the liking rating for the ninth baseline stimulus (in milliseconds).

[b]Liking ratings are reported as the difference between the liking ratings for the tenth target stimulus and the ninth baseline stimulus.

2.4.3.1 Response delay. The ANOVA revealed a significant main effect of the degree of deviance, $F(2,554) = 41.30$, $p < .001$, $\eta_p^2 = .13$, a significant main effect of expectation constraints, $F(2,554) = 7.15$, $p = .001$, $\eta_p^2 = .03$, and a significant interaction between the two factors, $F(4,554) = 3.45$, $p = .008$, $\eta_p^2 = .02$. Simple main effects analyses indicated that the effect of the degree of deviance was significant for all three expectation constraints conditions (all Fs ≥ 5.86, all ps $\leq .003$). Independent-samples t-tests revealed that the response delay was stronger for medium-deviant stimuli than for non-deviant stimuli (all ts ≥ 3.15, all ps $< .002$), and it was also stronger for highly- than for non-deviant stimuli (all ts ≥ 3.48, all ps $< .001$). The response delay difference between highly- and medium-deviant stimuli was only significant for highly-constrained expectations, $t(122) = 2.57$, $p = .011$, 95% $CI_{difference} = [241, 1861]$, $d = 0.46$, whereas there were no significant differences for moderately-constrained expectations, $t(123) = 1.35$, $p = .18$, or lowly-constrained expectations, $t(126) = 0.10$, $p = .92$.

The simple main effect of expectation constraints was only significant for highly-deviant stimuli, $F(2,554) = 11.02$, $p < .001$, $\eta_p^2 = .04$. The response delay was stronger for highly-constrained expectations compared to moderately-constrained expectations, $t(94.16) = 3.57$, $p = .001$, 95% $CI_{difference} = [586, 2057]$, $d = 0.64$, and compared to lowly-constrained expectations, $t(93.35) = 3.43$, $p = .001$, 95% $CI_{difference} = [533, 2001]$, $d = 0.62$. However, the response delay difference between moderately- and lowly-constrained expectations was not significant, $t(125) = -.21$, $p = .83$. We did not obtain significant simple main effects of expectation constraints for medium-deviant stimuli, $F(2,554) = 1.95$, $p = .14$, and for non-deviant stimuli, $F(2,554) = 0.97$, $p = .38$.

2.4.3.2 Liking ratings. We neither found a significant main effect of the degree of deviance, $F(2,554) = 1.81$, $p = .17$, nor of expectation constraints, $F(2,554) = 0.62$, $p = .54$, but crucially a significant interaction between the two factors, $F(4,554) = 2.71$, $p = .030$, $\eta_p^2 = .02$. Simple main effects analyses showed that the main effect of the degree of deviance emerged for lowly-constrained expectations, $F(2,554) = 3.77$, $p = .024$, $\eta_p^2 = .01$, and for highly-constrained expectations, $F(2,554) = 3.45$, $p = .033$, $\eta_p^2 = .01$. However, the only difference that remained significant after applying Bonferroni-correction was the difference between the liking ratings for highly- versus medium-deviant stimuli for lowly-constrained expectations, with highly-deviant stimuli being liked relatively less, $t(126) = -2.52$, $p = .013$, 95% $CI_{difference} = [-2.70, -0.33]$, $d = -0.45$. The simple main effect of the degree of deviance did not reach significance for moderately-constrained expectations, $F(2,554) = .03$, $p = .97$.

The analyses did not reveal a significant simple main effect of expectation constraints for any of the degree-of-deviance conditions (all Fs ≤ 2.96, all ps $> .05$).

2.4.3.3 Surprise ratings. The ANOVA showed a significant main effect of the degree of deviance, $F(2,569) = 42.25$, $p < .001$, $\eta_p^2 = .13$, and a significant main effect of expectation constraints, $F(2,542) = 6.39$, $p = .002$, $\eta_p^2 = .02$. The interaction between the degree of deviance and expectation constraints did not reach significance, $F(4,569) = 1.61$, $p = .17$. *T*-tests indicated that medium-deviant stimuli were rated as eliciting stronger feelings of surprise than non-deviant stimuli, $t(375.01) = 6.94$, $p < .001$, 95% $CI_{difference} = [1.61, 2.89]$, $d = 0.71$, and also highly-deviant stimuli received higher surprise ratings than non-deviant stimuli, $t(382) = 8.94$, $p < .001$, 95% $CI_{difference} = [2.12, 3.46]$, $d = 0.91$. The rating difference between highly- and medium-deviant stimuli was not significant, $t(389) = 1.74$. $p = .08$.

With a view to the main effect of expectation constraints, surprise ratings were higher for highly-constrained expectations compared to lowly-constrained expectations, $t(384) = 3.34$, $p = .001$, 95% $CI_{difference} = [0.47, 1.81]$, $d = 0.34$. After Bonferroni-correcting for multiple

comparisons, we did not observe significant rating differences between highly- and moderately-constrained expectations, $t(387) = 2.07$, $p = .04$, nor between moderately- and lowly-constrained expectations, $t(377.16) = 1.19$, $p = .23$.

2.4.3.4 Expectancy ratings. We found a significant main effect of the degree of deviance, $F(2,568) = 16.31$, $p < .001$, $\eta_p^2 = .05$, and a significant main effect of expectation constraints, $F(2,568) = 6.63$, $p = .001$, $\eta_p^2 = .02$, whereas the interaction between the two factors was not significant, $F(4,568) = 1.31$, $p = .27$. *T*-tests revealed that non-deviant stimuli were rated as more expected than medium-deviant stimuli, $t(379) = 4.62$, $p < .001$, 95% $CI_{difference}$ = [0.86, 2.13], $d = 0.47$, and than highly-deviant stimuli, $t(381) = 5.09$, $p < .001$, 95% $CI_{difference}$ = [1.04; 2.36], $d = 0.52$. The expectancy rating difference between medium- and highly-deviant stimuli was not significant. $t(388) = 0.66$, $p = .51$.

Follow-up *t*-tests on the main effect of expectation constraints showed that expectancy ratings were lower (i.e., unexpectedness was higher) for highly-constrained expectations compared to moderately-constrained expectations, $t(386) = -3.05$, $p = .002$, 95% $CI_{difference}$ = [-1.62, -0.35], $d = -0.31$, and compared to lowly-constrained expectations, $t(383) = -3.12$, $p = .002$, 95% $CI_{difference}$ = [-1.67, -0.38], $d = -0.32$. The expectancy rating difference between moderately- and lowly-constrained expectations was not significant, $t(379) = 0.12$, $p = .90$.

2.4.3.5 Correlational analyses. As Table 6 shows, all correlations were in the predicted direction except for the correlations between liking ratings and expectancy ratings, and between liking ratings and surprise ratings. Significant correlations were obtained between response delay and surprise ratings, and between surprise ratings and expectancy ratings. The overall association strength between the four surprise syndrome components was only low.

Table 6

Experiment 2: Pooled Pearson correlations among the four assessed surprise syndrome components. Statistically significant results are marked with an asterisk ($p \leq .008$, Bonferroni-corrected).

Syndrome component	1	2	3	4
1. Response delay	-			
2. Liking ratings	-.01	-		
3. Surprise ratings	.22*	.02	-	
4. Expectancy ratings	-.03	-.03	-.21*	-

2.4.4 Discussion

Experiment 2 aimed to refine and corroborate the preliminary evidence obtained in Experiment 1. Opposed to the pattern of a graded response delay in Experiment 1, Experiment 2 revealed a rather dichotomous behavioral response pattern, with a significantly stronger response delay for deviant stimuli compared to non-deviant stimuli, but no consistent differences between medium and high degrees of deviance. We will come back to discussing the overall implications of these diverging findings in the General Discussion. The current findings did not reveal evidence for systematic effects of the degree of deviance or expectation constraints on the affective surprise responses, suggesting that the pattern observed in Experiment 1 was indeed mainly driven by characteristics inherent to the stimuli employed. The subjective surprise ratings again reflected a rather dichotomous experiential feeling of being surprised (for both medium- and highly-deviant stimuli) or being not surprised (for non-deviant stimuli). Regarding the impact of expectation constraints, only the surprise rating difference between the two extreme categories (highly- vs. lowly-constrained expectations) reached significance, with the moderate constraints condition lying in between. Regarding expectancy ratings, we found a dichotomous cognitive appraisal of the event as being either

expected (for non-deviant stimuli) or unexpected (for both medium- and highly-deviant stimuli). Expectancy ratings were overall lower for highly-constrained expectations, but there were no further differences between moderately- and lowly-constrained expectations, which contrasts with our assumptions and with the results obtained in the pretest. Correlational analyses again revealed an overall only low association between the different surprise components.

2.5 General Discussion

Merging recent surprise theories renders the prediction that surprise is a function of how strong an event deviates from what was expected and of how easily it can be integrated with the constraints of an expectation. However, to our best knowledge, no previous research has systematically investigated the influences of both these two factors on multiple surprise responses. The present experiments therefore manipulated the degree of deviance of an event and, orthogonally to that, the constraints of a current expectation, while measuring the behavioral, affective, experiential and cognitive indicators of surprise. We predicted an increase of surprise responses with increasing degrees of stimulus deviance and increasing expectation constraints. First, we will summarize the results and discuss their theoretical implications, separately for each of the four surprise measures, and then continue with a discussion of the more general theoretical implications.

2.5.1 Summary of results – What's in a surprise?

2.5.1.1 Behavioral interruption. Realizing systematic degrees of deviance within one experiment, the results obtained in Experiment 1 empirically bolster the claim that behavioral surprise responses are susceptible to different degrees of deviance, which is in line with earlier findings (see Reisenzein et al., 2006). Conversely, Experiment 2 reveals dichotomous response patterns, with stronger response delays for both medium- and highly-deviant stimuli compared to non-deviant stimuli, but no significant differences between a medium and a high degree of

deviance. In the current research, behavioral surprise responses were operationalized via the time it takes participants to indicate their degree of liking for a (more or less) deviant stimulus compared to the preceding non-deviant stimulus. Thus, the observed response delay comprises not only the initial interruption that is directly triggered by the mere perception of the schema-discrepant event, but also an advanced cognitive processing and affective analysis (see also Horstmann, 2005; Meyer et al., 1997). Given the entanglement of these two processes, we tentatively speculate that the findings obtained in Experiment 1 may be an artifact of stimulus effects, with the more familiar non-word and picture stimuli requiring overall longer time to be evaluated compared to the relatively unfamiliar Chinese ideographs[6]. The evidence obtained in the more controlled Experiment 2 rather suggests that behavioral responses do not differentiate between different degrees of deviance. This contrasts previous research from Reisenzein et al. (2006) who report an effect of the extremity of stimulus change on response times. We think that this divergence might be due to methodical differences: Whereas Reisenzein et al. (2006) employed an unrelated performance task for reaction time measurements, our measurement of the response delay requires an additional evaluation of the degree of stimulus liking which might be confounded with affective processes. Combining the current degree-of-deviance manipulation with such a basic performance task in future studies will surely contribute to addressing this limitation and to gaining a refined understanding of the behavioral responses to surprise.

Across all experiments, we did not find any systematic effects of expectation constraints on the behavioral responses, implying that the broadness of explicit prior top-down expectations does not influence the ease and speed of surprise processing. This runs contrary to previous findings from Niepel (2001) who reports response time reductions for announced stimulus changes. As Niepel (2001) provided information on both *that* a change will take place and *when*

[6] We thank an anonymous reviewer for this thought.

it will take place, the pattern divergence may be explained by a lack of temporal anticipatability in our experiments. However, since we instructed participants that nine homogeneous stimuli and one surprising item will appear, passing nine homogenous trials should equally allow for anticipating that the surprise will take place in the tenth trial.

2.5.1.2 Affective responses. After controlling for stimulus effects, we did not find any systematic impact of the present manipulations on the affective surprise responses. This finding informs current theorizing, since recently there is a debate on whether spontaneous (negative) affect should be considered as a component of the surprise syndrome. While some influential authors do not deem affect as essential in surprise (for a recent comprehensive discussion, see Reisenzein et al., 2019), others focus strongly on the affective outcomes of surprise (Noordewier & Breugelmans, 2013; Noordewier et al., 2016), while still others even see negative affect as being the causal core mechanism of surprise, actually evoking all other surprise components (Topolinski & Strack, 2015; see also Proulx et al., 2012). The present data favor the former accounts that give little importance to affective components *inherent* to surprise. Our findings rather suggest that surprise has no specific valence, and that affective responses to surprising events are driven by the context-specific hedonic properties of the event itself. On the other hand, the liking ratings that participants had to render in the current experiments most probably do not reflect spontaneous affective reactions, but retrospective affect estimates that are contaminated by cognitive processing. Follow-up studies might therefore shed more light on the initial valence of surprise detached from advanced cognitive appraisals by implementing more subtle and direct measures that permit the implicit measurement of immediate and "on-line" affective responses, such as facial electromyography (see also Levy et al., 2018; Topolinski & Strack, 2015).

2.5.1.3 Experiential surprise feeling. The subjective surprise ratings reflected a rather dichotomous experiential feeling of being surprised (for both medium- and highly-deviant

stimuli) or being not surprised (for non-deviant stimuli). Whereas we found inconsistent effects of expectation constraints in Experiment 1, the results from Experiment 2 are in line with our assumption of increasing surprise ratings with increasing expectation constraints. However, we did not observe significant differences between all three expectation constraints conditions but only between the extreme categories (highly- vs. lowly-constrained).

When interpreting these results, it should be kept in mind that we measured the experiential feeling of surprise via retrospective, explicit surprise ratings, and after asking participants to indicate how much they like the surprising stimulus. Some authors would argue that, from a sense-making perspective, such ratings of surprise do not solely reflect experiential responses but rather the cognitive assessment of how easily the event can be explained and integrated with an existing representation (e.g., Foster & Keane, 2015; Maguire & Keane, 2006; Maguire et al., 2011). Therefore, it remains unclear whether we truly assessed the pure, initial feeling of surprise, or whether our findings instead reflect some residual "muddled measurements" (Noordewier et al., 2016, p. 138) of the primary surprise feeling and advanced sense-making processes (see also Pezzo, 2003; Schützwohl, 1998). Given the only low association between ratings of surprise and ratings of expectancy, we assume that we have measured two distinct concepts that are affected by cognitive appraisals to at least some markedly different degrees. However, to further restrict the impact of cognitive elaboration on experiential judgments, assessing the feeling of surprise directly after the deviant stimulus without intervening questions or imposing a time restriction for ratings (see Müller & Stahlberg, 2007, for a similar approach) may present a useful solution for future studies.

2.5.1.4 Cognitive appraisal of unexpectedness. Paralleling the pattern observed for the subjective surprise ratings, both experiments revealed a dichotomous cognitive appraisal of the event as being either expected (for non-deviant stimuli) or unexpected (for both medium- and highly-deviant stimuli). This implies that cognitive appraisals of unexpectedness are not

susceptible to different degrees of surprising-ness but rather use categorical representations ("Was it expected or not?").

The only impact of expectation constraints was found in Experiment 2, with significantly stronger unexpectedness appraisals for highly-constrained expectations. That we did not find a significant interaction between the degree of deviance and expectation constraints is interesting insofar as announcing a surprising stimulus obviously does *not* necessarily induce a mindset of "expecting the unexpected", which would have manifested in high expectancy ratings for deviant stimuli, but lower expectancy ratings for non-deviant stimuli. Thus, expecting the unexpected does neither consistently prevent, nor attenuate the strength of the surprise responses – we will come back to discussing the broader implications of this finding in Chapter 2.5.2.

2.5.1.5 Correlations among the surprise indicators. Providing one of the biggest correlative data sets in the surprise literature, we found only moderate to low intercorrelations between the different surprise components, which is in accordance with previous assessments of the strength of association among the multiple surprise syndrome components (Reisenzein, 2000a). It is noteworthy that also the correlation between ratings of surprise and expectancy which have been previously found to correlate strongly (Reisenzein, 2000a; Stiensmeier-Pelster et al., 1995; see also Reisenzein et al., 2019) was only $r \leq -.21$[7]. We cannot exclude that the low correlation coefficients currently obtained at least partly reflect methodical artifacts, resulting from the use of a between-subjects design (see Reisenzein, 2000a, for a thorough examination of this issue; see also Ruch, 1995). However, given our large sample sizes of N = 679 (Experiment 1) and N = 578 (Experiment 2), and given that stable correlation estimates emerge from approximately N = 250 (Schönbrodt & Perugini, 2013), we are confident that our findings present a valuable contribution to the present literature.

[7] Corrected statistics; the original version of the published manuscript mistakenly reports $r \leq -.25$.

2.5.2. General theoretical implications – What drives a surprise?

The body of evidence that was obtained in the present experiments points – except for the affective responses – towards relatively homogenous response patterns, with dichotomous responses that differ between "deviant" and "not deviant", but not between different grades of deviance. Opposed to our assumptions, we found only sporadic and inconsistent evidence on an impact of expectation constraints.

In a nutshell, the current results hence suggest that surprise is definitely about deviance, but not necessarily about the *degree* of deviance, and probably not about the difficulty of integrating an event with the constraints of an expectation. However, this may not necessarily imply that expectation constraints do not play any role at all, but rather that surprise is not primarily about top-down expectations evoked prior to the experimental task. We conducted the experiments with the crucial assumption that we would explicitly and actively induce different expectations by means of the different instructions at the beginning of the task. Our pretest provides convincing evidence that these different instructions indeed activated expectations with different constraints when asking participants to simulate the task at hand. However, given that participants encountered a sequence of nine familiarization trials in the main experiment, and given that we assessed expectancy ratings only after the surprising event, we cannot be certain that the initial expectation remained the dominant one. It could just as well be that passive expectation formations during the familiarization trials simply "overwrote" the initial explicit expectation, or that it was only after inquiring ratings that participants became post-hoc aware of what they had expected.

Our data rather point towards a primacy of passive expectation formation and suggest that expectations form up and alter dynamically over the course of a task. This is in line with recent neuro-cognitive theories on *predictive coding* (e.g., Clark, 2013; Feldman & Friston, 2010; Friston, 2010; Friston & Kiebel, 2009; Rao & Ballard, 1999) that claim a continuous

perception-based bottom-up adjustment of mental models: Instead of referring each stimulus back to the initial expectation, it appears that people rather calibrate and "fine-tune" this expectation during the task to permit an even better anticipation of the next stimulus (see also Hohwy, 2012; Horstmann, 2006; Itti & Baldi, 2009; Rumelhart, & Norman, 1978; Schützwohl, 1998). This is, across all conditions, seeing nine stimuli of the same type makes us refine and narrow our expectation to exactly this stimulus type, thus leading to the expectation that the next stimulus is of this type as well. Since the current experiments were not specially designed to determine whether and, if so, at what point in time experiences become more important than explicit instructions (and thus implicit expectation fine-tuning becomes the driving mechanism), this conclusion remains on a preliminary theoretical dimension. Future research might further investigate the relational dynamics between instructions and experience by combining our degree-of-deviance manipulation with a manipulation of the frequency of expectation activations prior to the surprising event (see Schützwohl, 1998), or by maintaining the salience of the initial expectation throughout the experiment.

Assuming an expectation fine-tuning over the course of the task seems also reasonable from a cognitive-evolutionary perspective that conceptualizes surprise as an adaptation-enabling mechanism (see Meyer et al., 1991, 1997; Reisenzein et al., 2019): In order to minimize prediction errors and maximize anticipation success, the organism needs to continually monitor its environment and to flexibly recalibrate its mental situational model. From this point of view, it also appears plausible that *any* signal that exceeds a certain discrepancy threshold should elicit the full immediate behavioral and experiential surprise mechanism, which is in line with the observed dichotomous response patterns.

While a lacking impact of expectation constraints seems most intuitive for the behavioral responses – knowing, for example, that a change will appear may probably not prevent you from being initially interrupted –, it is all the more interesting that we do not

observe an effect of active surprise anticipation (low expectation constraints) on ratings of surprise or on ratings of expectancy. This contrasts with the assumptions of sense-making approaches according to which self-reports of surprise depend on the ease of integrating an event with an existing representation (e.g., Foster & Keane, 2015; Maguire et al., 2011; Maguire & Keane, 2006). From a sense-making perspective, partial explanations should facilitate the resolution of discrepancies. Thus, a (medium- or highly-)deviant stimulus should be easier to explain than a non-deviant stimulus when one actively expected a surprise to appear, whereas the same (medium- or highly-)deviant stimulus should be harder to explain when one did not expect such a surprise. However, that we did not observe a significant interaction between the degree of deviance and expectation constraints on ratings of surprise or expectancy in the current experiments rather suggests that people do *not* use their prior explicit knowledge for calculating their level of surprise. Future research may carefully test this preliminary implication by investigating the impact of explanatory difficulty and enabling information more closely (see also Maguire et al., 2011) to see whether integration-facilitating cues indeed foster surprise resolution.

With a view to the cognitive appraisal of unexpectedness, it might come as a potential limitation of our design that expectancy ratings were assessed only post-hoc and could thus be influenced by hindsight bias[8]. Since the current literature predicts hindsight bias only for successfully integrated events, this would lead us to expect partly the same pattern as observed: relatively high expectancy ratings (and low surprise ratings) for non-deviant events that can be easily made sense of, and lower expectancy ratings (and high surprise ratings) for both medium- and highly-deviant events that cannot be fully resolved. However, we would have also expected a pattern reversal for the lowly-constrained condition that explicitly announced a surprise: If expectancy ratings indeed reflected an "I should have known it all along" phenomenon (Pezzo,

[8] We thank an anonymous reviewer for pointing us to this thought.

2003), expectancy ratings should have been high for deviant stimuli in this condition. We assume that the divergence between our findings and previous sense-making evidence results from methodical differences in the paradigms employed: Whereas most sense-making research builds on hypothetical scenario situations, the present research elicited surprise in-vivo, thus increasing the realism of the current results.

At the outset of this paper we have argued that surprise describes the reaction to an unexpected event; yet, concluding, our findings put this prominent assumption into question. Firstly, the correlation between surprise and expectancy ratings was astonishingly low in the present data, $r \leq -.21$[9], which was even lower than in previous work (Reisenzein, 2000a; Stiensmeier-Pelster et al., 1995). Secondly, the observation that even informing participants of an upcoming surprise does neither prevent, nor significantly attenuate the surprise responses indicates that stimuli do not necessarily need to be explicitly unexpected to be surprising, but that surprise rather results from the conflict between implicitly formed and dynamically fine-tuned expectations and perceived events (see also Retell et al., 2016, for a similar discussion in the context of visual search). This implies that talking about surprise as an outcome of unexpectedness requires a clear definition of what is meant by an "expectation" first.

With a view to the impact of the degree of deviance, we deem it important to cautiously qualify our conclusion of a dichotomy of surprise responses to the current experimental setting and choice of stimuli. Since the descriptive results patterns indeed point towards increasing surprise responses with increasing degrees of deviance, we cannot preclude that a set of more engaging and meaningful stimuli than the highly controlled, but also highly impoverished stimuli in Experiment 2 might intensify surprise reactions, causing differences that also reach statistical significance. For these reasons, the present paper cannot conclusively answer the question of whether surprise is an all-or-nothing phenomenon or a graded experience. However,

[9] Corrected statistics; the original version of the published manuscript mistakenly reports $r \leq -.25$.

we are confident that the current experiments present an important first step in researching the graded-ness of multiple surprise responses in a controlled experimental setting.

Another important potential shortcoming of the present experiments is their reliance on Western, young and highly-educated student samples with a disproportionately large number of female participants. Since this presents a common deficiency among psychological research and since also previous evidence builds on samples that are composed in a similar manner, our findings are certainly integrable into the current empirical and theoretical context. Given the basic cognitive mechanisms of surprise that we investigated, we would also not expect these observations to be different for more heterogeneous and inclusive samples. However, the extendibility of our results to different populations with varying cultural and social backgrounds, age, or gender distributions remains only speculation. Replicating the present experiments with more diverse samples may thus provide another fruitful avenue for future research to ensure the generalizability of our findings.

2.6 Conclusion

To conclude, our findings imply that surprise presents a multi-faceted, complex phenomenon that results from the temporal interplay between perceptual input and the continuous fine-tuning of expectations. Instead of primarily reflecting the (strength of) violation of initial top-down expectations or the ease of integrating an event with an explicit representation, surprise rather seems to be the automatic outcome of implicit discrepancy detection. This underlines the need for more comprehensive theories and deeper research on the causal architecture of surprise that integrate existing approaches and empirical evidence into a unified perspective.

Within the scope of unpredictability, an investigation of the unpredicted presents, however, only one side of the coin, and the unpredictable remains to be focused. Since unpredictable events have obviously not happened yet, there are, in logical consequence, no

manifest responses to assess. Nonetheless, what can be measured are affective evaluations of encountering unpredictable situations. On these backgrounds, Chapter 3 intends to explore the valence of the unpredictable by deriving the value that people are willing to forgo to ensure predictable (vs. unpredictable) social interactions.

2.7 Appendix

2.7.1 Appendix A: Supplementary materials

De-identified raw and aggregated data files and supplementary analyses are available at the Open Science Framework: https://osf.io/68du7/

2.7.2 Appendix B: Condition assignments in Experiments 1 and 2

Table 7

Experiment 1: Sample characteristics and baseline condition assignments.

		Expectation constraints		
		High	Moderate	Low
Sample characteristics		M_{age} = 23, SD_{age} = 4; 157 female, 70 male missing age information from *n* = 3	M_{age} = 23, SD_{age} = 6; 154 female, 62 male, 1 gender diverse missing demographic data from *n* = 2; missing age information from *n* = 1	M_{age} = 22, SD_{age} = 5; 159 female, 68 male, 3 gender diverse missing demographic data from *n* = 3; missing age information from *n* = 5
Condition assignments				
No deviance	Non-words	*n* = 37	*n* = 38	*n* = 40
	Pictures	*n* = 42	*n* = 37	*n* = 40
Medium deviance	Non-words	*n* = 38	*n* = 36	*n* = 39
	Pictures	*n* = 34	*n* = 37	*n* = 38
High deviance	Non-words	*n* = 38	*n* = 35	*n* = 39
	Pictures	*n* = 38	*n* = 36	*n* = 37

Table 8

Experiment 2: Baseline condition assignments.

		Expectation constraints		
		High	Moderate	Low
	Baseline cond.			
No deviance	Green numbers	*n* = 17	*n* = 17	*n* = 19
	Blue numbers	*n* = 14	*n* = 16	*n* = 14
	Green letters	*n* = 15	*n* = 15	*n* = 13
	Blue letters	*n* = 17	*n* = 16	*n* = 14
Medium deviance	Green numbers	*n* = 17	*n* = 18	*n* = 19
	Blue numbers	*n* = 14	*n* = 14	*n* = 17
	Green letters	*n* = 15	*n* = 15	*n* = 15
	Blue letters	*n* = 20	*n* = 15	*n* = 15
High deviance	Green numbers	*n* = 16	*n* = 16	*n* = 18
	Blue numbers	*n* = 14	*n* = 16	*n* = 17
	Green letters	*n* = 15	*n* = 15	*n* = 15
	Blue letters	*n* = 23	*n* = 19	*n* = 13

2.7.3 Appendix C: Stimuli used in Experiment 1

Verbal stimuli were taken from a pool of non-words with unsystematic transitions of consonantal articulation from Topolinski et al. (2014).

DERUBA, DIGABO, GABULO, GEBISA, KABIDU, KOBUNE, LIKOBE, NARUBO, NEGIBA, RUBALO

Pictorial stimuli were taken from the IPNP online database (Szekely et al., 2004). Full materials can be downloaded from https://crl.ucsd.edu/experiments/ipnp/dataquery/querymini.php.

obj022bag, obj042bench, obj045bird, obj056box, obj081car, obj099clock, obj180glass, obj234lamp, obj236leaf, obj263mirror

Chinese ideographs were taken from an item pool from Topolinski and Strack (2009).

普 助 词 代

副 动 名 氏

池 连

Chapter 3 – The Price of Predictability – Estimating Inconsistency Premiums in Social Interactions

Abstract

For financial decision-making, people trade off the expected value (return) and the variance (risk) of an option, preferring higher returns to lower ones and lower risks to higher ones. To make a decision-maker indifferent between a risky and risk-free option, the expected value of the risky option must exceed the value of the risk-free option by a certain amount – the *risk premium*. Previous psychological research suggests that similar to risk aversion, people dislike inconsistency in an interaction partner's behavior. In seven experiments (total N = 2,261) we pitted this inconsistency aversion against the expected returns from interacting with an inconsistent partner. We identified the additional expected return of interacting with an inconsistent partner that must be granted to make decision-makers prefer a more profitable, but inconsistent partner to a consistent, but less profitable one. We locate this *inconsistency premium* at around 31% of the expected value of the risk-free option.

3.1 Introduction

People like predictability. We watch tomorrow's weather report, consult opinion portals to assess whether the new restaurant around the corner will suit our taste, observe share price to predict stock performance, and condense the world into logical axioms that aim at forecasting the course of events. Reducing uncertainty and increasing predictability has been discussed as a fundamental human need (Heider, 1958; Hogg, 2000; Kagan, 1972), with the pleasure of predictability deriving from the perceived ability to anticipate and control our environment. Whereas predictability facilitates attentional orienting, processing, and performance (e.g.,

Alink, Schwiedrzik, Kohler, Singer, & Muckli, 2010; Coull & Nobre, 1998; Posner, Snyder, & Davidson, 1980) and is processed as rewarding by the brain (e.g., Braem & Trapp, 2019; Trapp et al., 2015), unpredictability is experienced as aversive (e.g., Heine et al., 2006; Proulx et al., 2012; Schubert, Körner, Lindau, Strack, & Topolinski, 2017; Topolinski & Strack, 2015) and increases stress and physiological arousal (e.g., de Berker et al., 2016; Herry et al., 2007; Jackson, Nelson, & Proudfit, 2015; Mendes, Blascovich, Hunter, Lickel, & Jost, 2007; Peters et al., 2017).

The benefits of predictability are thus evident. However, with a world of dynamic change, uncertainty is a fundamental feature of reality, and its sources are manifold. Some uncertainties can be described as *stochastic* which means that a process sometimes leads to outcome A and sometimes to outcome B while "nature decides" which outcome will occur. For instance, if you cast a dice, "nature" will decide which number of eyes will come up. To be sure, the exact characteristics of the process (shape of the dice, power of the throw, etc.) eventually determine the outcome, but assuming a random process executed by nature is a feasible theoretical approach to such phenomena. In contrast, for many private and public decisions, uncertainty is *strategic*. That is, the uncertain outcome is not determined by a random process executed by nature, but by a person making different decisions (Brandenburger, 1996; see also Heinemann, Nagel, & Ockenfels, 2009; Knight, 1921). While stochastic uncertainty can only be dealt with in terms of probability distributions, strategic uncertainty may be mitigated by considering the beliefs, intentions and preferences of another decision-maker (see Camerer, 2003). However, decision-makers facing strategic uncertainty may also have certain preferences regarding the uncertainty about another person's decision.

In general, when making decisions, people strive to maximize returns and to minimize risk (see Coombs, 1975). In finance, these concepts are referred to as the *expected value* (return) and the *variance* (risk) of an investment (e.g., Markowitz, 1952; Sharpe, 1964). Human

preferences are assumed to be organized in such a way that higher expected values are preferred to lower ones, while lower variances are preferred to higher ones (see Arrow, 1965; Bernoulli, 1738/1954; Pratt, 1964). Given two options with equal expected values but different variances, most people thus act risk-averse and prefer the option that shows the lower variance. Similarly, given equal variances but different expected values, one should prefer the option with the higher expected value. However, because most financial options that we face in our everyday lives differ in both their expected values and variances, decision-making is based on a *risk-return trade-off*: Reducing risk might come along with lower returns, and increasing returns might come along with higher risk (Merton, 1980; see also Ghysels, Santa-Clara, & Valkanov, 2005). Therefore, to compensate a decision-maker for higher risks, the returns must include an additional *risk premium*. More specifically, the risk premium is defined as the amount by which the expected value of the risky option exceeds the value of the risk-free option while the decision-maker is indifferent between both options (Merton, 1980; Pratt, 1964).

For example, imagine you were offered the choice between receiving $5 for sure or flipping a coin and receiving $10 if it comes up heads, but nothing if it comes up tails. Objectively, in this scenario, both options have the same expected value of $5. However, as a risk-averse decision-maker, you would probably prefer the certain outcome to the risky gamble. But what amount of money would we need to add to the expected value of the coin flip to make you indifferent between the risky and the risk-free option? Or, put differently, what is the additional return of the coin flip that must be exceeded to make you just prefer the risk? If increasing the expected value of the coin flip to $6 would make you indifferent between the certain payoff and the risky gamble, then the risk premium would be $1, or 20% of the expected value of the certain option – and any expected value of the coin flip that exceeds these $6 would probably make you prefer taking the risk.

From a social psychological perspective, variance in a person's behavior can be described in terms of consistency. Thus, because only consistent behavior allows predictability, people might – similar to risk aversion under stochastic uncertainty – exhibit inconsistency aversion under strategic uncertainty and favor consistent interaction partners who show no variance in their behavior to those displaying high variance by behaving inconsistently. However, whereas the preference for predictability is well-researched in the psychological literature, astonishingly little attention has been paid so far to the psychological trade-off between risk and return when facing strategic uncertainty. To our best knowledge, no previous research has applied the concept of a risk premium to social interactions, examining whether and what monetary value people are willing to forgo to interact with a consistent, but only moderately profitable partner instead of an inconsistent, yet more profitable one. To fill the present research gap, we systematically addressed this question within seven experiments.

3.2 Aim and design of the present research

The present research experimentally benchmarks the risk-return trade-off in social interactions under strategic uncertainty. Pitting inconsistency aversion against the expected returns from an interaction, we aim at identifying the additional expected value of an interaction that is necessary to make decision-makers indifferent between an inconsistent interaction partner and a consistent one. We will term this difference in the expected values of interacting with a consistent and an inconsistent partner that makes a decision-maker indifferent between the two partners the *inconsistency premium*. Crucially, this approach allows us to estimate the tipping point of preferences for consistency versus expected returns. So, for a given level of behavioral inconsistency, which additional expected return of the inconsistent interaction must be granted to make decision-makers prefer a more profitable, but inconsistent interaction partner to a consistent, but less profitable one? Or, turning this logic around, how much of their

potential returns are decision-makers willing to forgo to avoid uncertainty and to ensure consistency in social interactions?

We estimated the magnitude of the inconsistency premium in a set of seven experiments (total N = 2,261). Crucially, in each experiment, we systematically manipulated the expected value of interacting with the inconsistent partner across conditions while keeping the expected value of interacting with the consistent partner constant. While Experiments 1–3 probed the psychological value of consistency in an organizational workplace setting by exploring the effects of behavioral (in)consistency and expected return on collaboration preferences, Experiments 4–7 extended these findings to social interactions in an economic game. As dependent variables, we assessed both stated preferences (measured by explicit preference ratings for the consistent and inconsistent interaction partner; Experiments 1, 3–7) and revealed preferences (measured by the behavioral choice between the consistent vs. inconsistent interaction partner; Experiments 2, 5–7). Importantly, participants in Experiments 5–7 played an incentivized round of an economic game with the player of their choice. That is, they actually had to pay a price (in terms of the expected value) if they chose a consistent, but less profitable interaction partner. Deploying both self-report and behavioral measures as two hitherto coexisting measurement traditions (for a recent discussion, see Frey, Predroni, Mata, Rieskamp, & Hertwig, 2017; Mata, Frey, Richter, Schupp, & Hertwig, 2018) thus provides a comprehensive view on inconsistency premiums in social interactions.

We are certainly not the first to explore the consequences of the risk-return trade-off for human decision-making. For example, previous research already approached such trade-offs from a neuro-economic perspective (e.g., Heinemann et al., 2009; Schmidt, Shupp, Walker, & Ostrom, 2003; see Preuschoff et al., 2006), compared risk preferences between human and computer interactions in the context of a Trust Game (Bohnet, Greig, Herrmann, & Zeckhauser, 2008; Bohnet & Zeckhauser, 2004; Fetchenhauer & Dunning, 2012) and investigated the impact

of different risk-return profiles on performance evaluations and wage payments (Barnes & Morgeson, 2007; Bodvarsson & Brastow, 1998; DeNisi & Stevens, 1986; Deutscher & Büschemann, 2016; Deutscher, Gürtler, Prinz, & Weimar, 2017; Newman, Krzystofiak, & Cardy, 1986). However, to our best knowledge, we are the first to transfer the concept of a risk-return trade-off to preferences in social interactions, aiming to quantify the psychological value of behavioral consistency.

Data treatment and a priori power-analysis. Due to the novelty of our research question, we could only estimate the effect size and assumed an average effect size of $d_z = 0.40$ for the rating difference between the inconsistent and the consistent interaction partner in a two-tailed paired-samples *t*-test. This would require $n = 52$ per cell to yield a power of 80%. To be able to detect medium effects of $h = 0.4$ with a power of 80% in a two-sided binomial test in Experiment 2, data from $n = 50$ participants per condition would be needed. Experiment 7 aimed to increase the internal validity of the inconsistency premium estimate and built on an assumed lower average effect size of $d_z = 0.25$, resulting in $n = 128$ per cell.

All experiments were conducted on Amazon's crowdsourcing platform *Mechanical Turk*. Statistical analyses were run after data of the full sample were collected. We report all manipulations, all measures, all exclusions of data and all preparatory steps prior to the analyses. Data from all experiments are available online at https://osf.io/ay6c3.

3.3 Experiment 1

Experiment 1 investigated the risk-return trade-off in an organizational workplace setting. Social interactions were operationalized by collaborations on a future job project. Participants were told to imagine that they would soon start a new project which is very important for their further career. Since they would need a second co-worker in their team, their task would be to indicate how much they would like to collaborate with two potential co-workers. As the only basis for their evaluations, they would be presented information on how

each person performed in ten previous projects. We manipulated both the variances of the co-workers' performances and the expected value of the inconsistent co-worker's performance. The consistent co-worker constantly exhibited a performance of 50% in each of the ten projects while the inconsistent co-worker exhibited a performance of either 10% or 90% in each of the ten projects. Crucially, across five between-conditions we increased how often the inconsistent co-worker exhibited a performance of 90% instead of 10%, thereby making the collaboration with the inconsistent co-worker more and more socially "profitable". Hence, whereas collaborating with the consistent co-worker always has an expected value of 50%, collaborating with an inconsistent co-worker who shows, for example, performances of 10% in four projects and performances of 90% in six projects, would yield a higher expected value of 58%.

We had a 2 (Co-worker behavior: consistent vs. inconsistent; within-participants) X 2 (Expected value of collaborating with the inconsistent co-worker: 50% vs. 58% vs. 66% vs. 74% vs. 82%) design. As the dependent variable, we measured participants' stated preferences on how much they would like to collaborate with each of the two co-workers on a new project. The preregistration of Experiment 1 can be assessed at https://osf.io/tcj4z/?view_only=b0a69ec898ac409890ac5753603be313.

3.3.1 Method

3.3.1.1 Participants. N = 250 participants (96 female, 154 male; M_{age} = 34, SD_{age} = 10) were recruited via Amazon's *Mechanical Turk* and received $0.5 for compensation. All participants indicated their informed consent.

3.3.1.2 Procedure. Participants were told to imagine that they work as an employee of a company and will soon start a new job project which is very important for their further career. However, they would need a second co-worker in their team and would therefore receive information on the previous performances of two potential co-workers before they decide with whom to collaborate. Participants then sequentially received information on the performances

of each co-worker in ten previous projects. All 20 information trials were presented in random order.

Inconsistency was manipulated within-participants by implementing one co-worker who consistently completed 50% of the tasks in each of the ten previous projects, whereas the other co-worker showed inconsistent performances and completed either 10% or 90% of the tasks. The expected value of collaborating with the inconsistent co-worker was manipulated between-participants by increasing the number of 90% versus 10% performances across five conditions (see Table 9, Appendix D). Participants were randomly assigned to one of the five conditions. To control for potential confounding of co-worker preferences with letter preferences, we additionally counterbalanced between participants the assignment of letter abbreviations (A vs. B) to co-worker behavior (consistent vs. inconsistent). Having studied the performances of the two co-workers, participants were asked to indicate how much they would like to collaborate with each of the two co-workers on their new project on a scale from 0 (*not at all*) to 10 (*very much*).

3.3.2 Results

A 2 (Co-worker behavior: consistent vs. inconsistent; within-participants) X 2 (Expected value of collaborating with the inconsistent co-worker: 50% vs. 58% vs. 66% vs. 74% vs. 82%) ANOVA did neither reveal a significant main effect of co-worker behavior, $F(1,245) = 1.40$, $p = .24$, nor a main effect of the expected value of collaborating with the inconsistent co-worker, $F(1,245) = 6.72$, $p = .08$. However, there was a highly significant interaction between the two factors, $F(4,245) = 12.22$, $p < .001$, $\eta_p^2 = .17$.

Pairwise comparisons between the preference ratings for the consistent and the inconsistent co-worker for each of the five between-conditions are depicted in Table 9 (see Appendix D). Given equal expected values of 50%, participants indeed preferred the consistent co-worker to the inconsistent one, thus implying that our paradigm was able to successfully

induce and measure psychological inconsistency aversion. The inconsistent co-worker was significantly preferred to the consistent co-worker with an expected value of 82%. There were no significant rating differences for the remaining conditions. Figure 3 depicts the overall collaboration preference pattern in terms of rating differences between the inconsistent and the consistent co-worker.

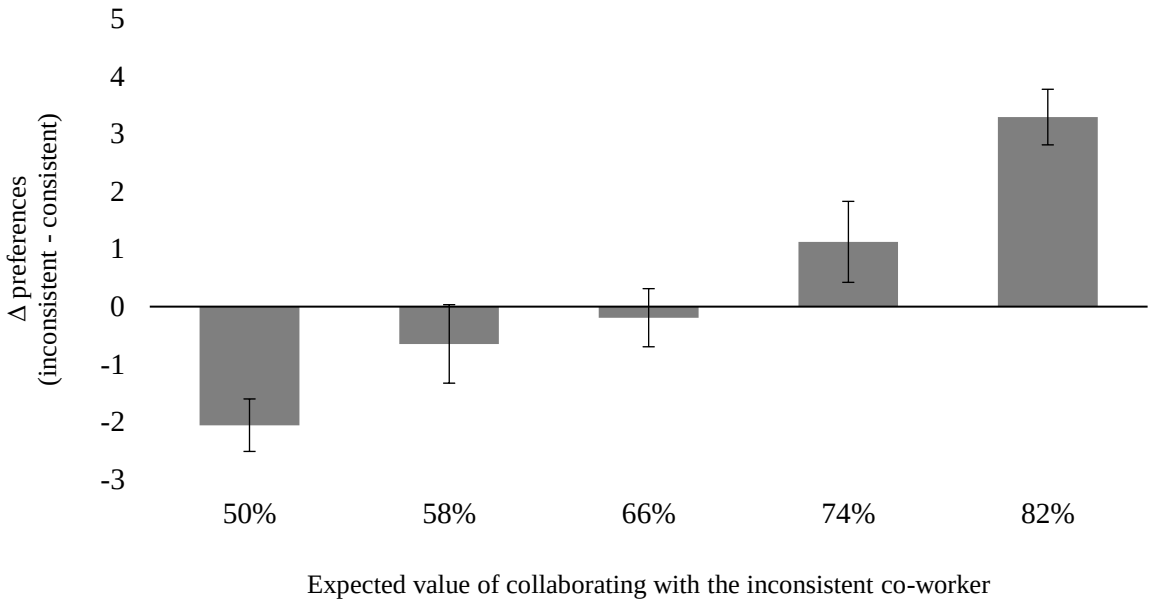

Figure 3. Collaboration preferences in Experiment 1. Bars depict the differences between the collaboration preference ratings for the inconsistent and the consistent co-worker. Positive values indicate a collaboration preference for the inconsistent co-worker. Negative values indicate a collaboration preference for the consistent co-worker. Error bars represent *SEM*s.

3.3.3 Discussion

Pitting consistent against inconsistent performances, we probed the minimum difference in expected values to make participants prefer an inconsistent, but more profitable co-worker to a consistent, yet less profitable one. Experiment 1 suggests a significant consistency preference for equal expected values, with preferences shifting in favor of the inconsistent co-

worker only when the expected value of collaborating with the inconsistent co-worker was increased to 82%. These findings imply an inconsistency premium of around 32%.[10]

However, our present inconsistency premium estimate builds on explicit co-worker evaluations, that is, on stated preferences that may – or may not – predict real decisions and thus the actual payoff that people are willing to forgo in favor of predictable interactions (e.g., Cummings, Harrison, & Rutström, 1995; Johannesson, Liljas, & Johansson, 1998; Lambooij et al., 2015). To obtain a more direct measure we therefore additionally assessed revealed preferences as indexed by behavioral choices between the consistent vs. inconsistent co-worker in a second experiment.

3.4 Experiment 2

To corroborate the evidence obtained in Experiment 1, Experiment 2 employed a binary choice paradigm to measure revealed preferences, asking participants to decide whether they would like to collaborate with either the consistent or the inconsistent co-worker on a future work project. The preregistration of Experiment 2 can be assessed at https://osf.io/dv3fj/?view_only=a7e0961066144ee2871c595f91fc0401.

3.4.1 Method

3.4.1.1 Participants. N = 251 participants (91 female, 160 male; M_{age} = 37, SD_{age} = 12) were recruited via Amazon's *Mechanical Turk* and received $0.5 for compensation. Informed consent was obtained from all participants.

[10] The inconsistency premium reflects the additional (relative) return that must be added to the expected return of interacting with a consistent partner to make participants indifferent between the consistent and the inconsistent partner. To obtain our estimate, we defined the consistent partner's expected value as the reference point of 100% and calculated the relative excess of the observed point of indifference. In case of multiple indifference conditions, the mean excess was used as an estimate.

3.4.1.2 Procedure. The procedure was the same as in Experiment 1 with the only difference that this time, participants were asked to choose one of the co-workers for an upcoming collaboration instead of rating them.

3.4.2 Results

A chi-square test for independence indicated a significant association between the expected value of collaborating with the inconsistent co-worker and co-worker choice, $\chi^2(4) = 11.80$, $p = .019$, $V = .36$. Separate binomial tests (against chance level of 50%) for each between-condition revealed that for equal expected values, participants significantly preferred the consistent co-worker. In contrast, in the highest expected value condition (82%), participants significantly preferred the inconsistent co-worker. There were no significant differences in the remaining conditions (see Appendix D, Table 10).

3.4.3 Discussion

Experiment 2 replicated the pattern that was obtained for stated preferences in Experiment 1. For equal expected values, participants exhibited a significant preference for consistency that was reversed only by the prospect of higher returns. This implies an inconsistency premium of around 32% for actual decisions about future collaboration partners.

In a next step, we aimed to investigate whether the inconsistency premium spills over to more general social evaluations.

3.5 Experiment 3

While previous research already points out an impact of predictability on the likeability of social groups (see Hinds, Carley, Krackhardt, & Wholey, 2000; Worthen, Coats, McGlynn, & Rossano, 2007), it remains hitherto unexplored whether and how people trade off behavioral inconsistency (i.e., variance) and expected value when they generally evaluate how much they like another person. Whereas the previously assessed preferences most likely reflect the trade-

off between participants' perceived personal risks and returns that arise from collaborating with the respective co-workers, liking evaluations do neither entail any risks, nor potential returns, and thus no risk-return trade-off in the common understanding. Thus, Experiment 3 investigated the downstream consequences of inconsistency and tested to what degree the inconsistency premium also manifests in the affective evaluation of a potential co-worker. The preregistration can be assessed at https://osf.io/smvj5/?view_only=8b9eb3c4a10341e79864a20b5e2b6cd1.

3.5.1 Method

3.5.1.1 Participants. N = 251 participants (118 female, 132 male, 1 other; M_{age} = 36, SD_{age} = 10) were recruited via Amazon's *Mechanical Turk* and received $0.5 for compensation. Informed consent was obtained from all participants.

3.5.1.2 Procedure. The procedure was similar to Experiments 1 and 2 with the crucial difference that this time, participants were asked to indicate how much they like the consistent and the inconsistent co-worker on a scale from 0 (*not at all*) to 10 (*very much*).

3.5.2 Results

A 2 (Co-worker behavior: consistent vs. inconsistent; within-participants) X 2 (Expected value of collaborating the inconsistent co-worker: 50% vs. 58% vs. 66% vs. 74% vs. 82%) ANOVA revealed a significant main effect of co-worker behavior, $F(1,246) = 16.42$, $p < .001$, $\eta_p^2 = .06$, no significant main effect of the expected value of collaborating with the inconsistent co-worker, $F(4,246) = 4.64$, $p = .21$, but crucially a significant interaction between the two factors, $F(4,246) = 13.71$, $p < .001$, $\eta_p^2 = .18$. The consistent co-worker was preferred to the inconsistent co-worker with equal expected values of 50%. Whereas we observed indifference between the two co-workers when the expected value of collaborating with the inconsistent co-worker was 58%, preferences flipped in favor of the inconsistent co-worker starting from an expected value of 66%. For more details, pairwise comparisons between the

liking ratings for the consistent and the inconsistent co-worker for each of the five between-conditions are depicted in Table 11 (see Appendix D). Figure 4 plots the overall preference pattern in terms of rating differences between the inconsistent and the consistent co-worker.

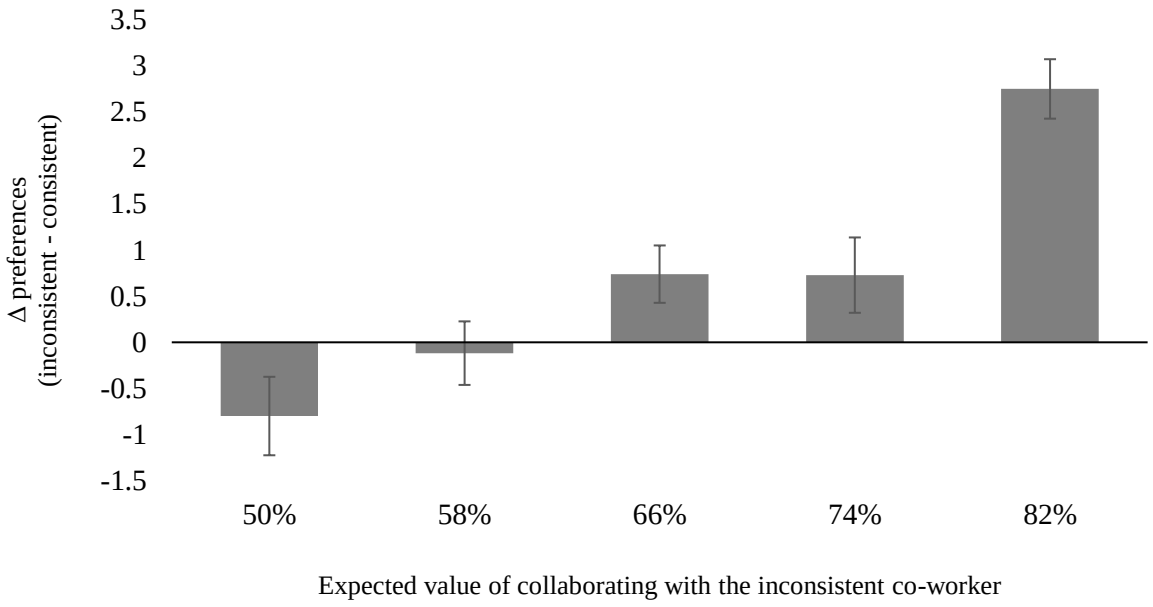

Figure 4. Co-worker liking ratings in Experiment 3. Bars depict the differences between the liking ratings for the inconsistent and the consistent co-worker. Positive values indicate a liking preference for the inconsistent co-worker. Negative values indicate a liking preference for the consistent co-worker. Error bars represent *SEM*s.

3.5.3 Discussion

Given that the principle of risk-return trade-offs originally stems from finance and refers to monetary investments, it is noteworthy that we also observed a psychological risk-return trade-off when assessing basic social evaluations of person liking. That is, affective evaluations mirror the idea of an inconsistency premium such that higher expected returns can increase the liking for people who behave inconsistently. We currently locate the magnitude of this affective inconsistency premium at around 16%.

In a next step, we aimed to test the robustness of our current inconsistency premium and examined participants' risk versus return preferences in a more abstract and formalized context.

3.6 Experiments 4a–e

While Experiments 1–3 investigated the inconsistency premium in an organizational workplace setting, Experiment 4 aimed to generalize the present findings to one of the most common currencies in our everyday lives: money. Therefore, in Experiments 4a–e, we used the *Ultimatum Game* (Güth, Schmittberger, & Schwarze, 1982) as our primary research paradigm (see Chang, Levinboim, & Maheswaran, 2012, for a related implementation). We decided to build on the Ultimatum Game because it is extremely well-researched (e.g., Gabay, Radua, Kempton, & Mehta, 2014; Güth & Kocher, 2014; Oosterbeek, Sloof, & van de Kuilen, 2004) and provides a paradigm where the personal payoff depends on the behavior of interaction partners, thus creating strategic uncertainty. Playing a sequence of games with different proposers allows to manipulate both the interactions' expected value (by varying the average size of the offers) and consistency (by varying the variance in offer size). Thereby we could implement two crucial proposer types: the consistent, moderately profitable proposer, and the inconsistent, but economically at least equally (or even more) profitable proposer.

In the following experiments, the social interactions were introduced as economic bargaining encounters. Participants played as responders who had to decide whether to accept or reject offers from different proposers who divide an amount of $10 between themselves and the responder (i.e., the participant). After ten rounds, they were then asked to indicate how much they would like to play further rounds with each of the proposers. Crucially, we manipulated within-participants the consistency of those proposers' offers. In detail, the consistent proposer offered a constant amount of $5 in all ten rounds while the inconsistent proposer offered either $0 or $10 in each round. Across five sub-experiments, we increased how often the inconsistent proposer offered $10 instead of $0, thereby making the interaction with the inconsistent

proposer more and more profitable. Thus, interacting with the consistent proposer always had an expected value of $5, whereas interacting with an inconsistent proposer offering for example nine times $10 and once $0 would have a higher expected value of $9.

3.6.1 Method

3.6.1.1 Participants. Participants were recruited via Amazon's *Mechanical Turk* and received $0.5 for compensation. Sample characteristics are displayed in Table 12 (see Appendix D). Data from participants who took part in more than one experiment were considered for analysis only for first participation, with all further data being excluded (Experiment 4d: $n = 1$; Experiment 4e: $n = 2$). We also excluded participants who took part more than once in one experiment (Experiment 4b: $n = 1$). Informed consent was obtained from all participants.

3.6.1.2 Procedure. We realized two crucial target proposers – the consistent and the inconsistent one – and intentionally implemented two further distractor proposers to make our manipulation less obvious. Participants were told that the experiment aimed at simulating financial bargaining situations. Instructions informed them that there would be two roles in the experiment – the role of the proposer and the role of the responder – and that they would always be assigned the role of the responder. They were instructed that they would play multiple bargaining rounds involving the four different proposers A, B, C, and D, who make an offer how to divide an amount of $10 between themselves and the responder (i.e., the participant). Their task would be to accept or reject an offer, with offer acceptance resulting in the payoff specified in the proposer's offer, and with offer rejection resulting in payoff cancellation for both the proposer and the responder. Participants then played a sequence of ten rounds with each of the four different proposers, thus making a total of 40 trials. Offers from all proposers were presented in random order. The offers of the consistent proposer and the two distractor proposers were kept constant across Experiments 4a–e. Whereas the consistent proposer offered $5 in each trial (expected value: $5), the first distractor proposer offered three times $4, four

times \$5, and three times \$6 (expected value: \$5), and the second distractor proposer offered five times \$1 and five times \$10 (expected value: \$5.5).[11] Crucially, we manipulated the expected value of interacting with the inconsistent proposer across experiments by varying the instances of \$0 vs. \$10 offers. To control for potential confounding of proposer preferences with letter preferences we counterbalanced between participants the assignment of letter abbreviations (A, B, C, or D) to proposer behavior (consistent, inconsistent, or distractor proposer). After the bargaining sequences, participants were asked to indicate how much they would like to play further bargaining rounds with each of the proposers (identified by the letter) on a scale from 0 (*not at all*) to 10 (*very much*).

3.6.2 Results

Pairwise comparisons between the stated preferences for the consistent and the inconsistent proposer for each experiment are shown in Table 12 (see Appendix D). The results indicate that with equal expected values of \$5, the consistent proposer was significantly preferred to the inconsistent proposer (Experiment 4a). This consistency preference was also evident when the expected value of interacting with the inconsistent proposer was increased to \$6 (Experiment 4b) and \$ 7 (Experiment 4c). With an expected value of \$8, participants were indifferent between the two proposers (Experiment 4d). A significant preference for the inconsistent proposer only emerged when the expected value was increased to \$9 (Experiment 4e). Figure 5 plots the overall preference pattern in terms of rating differences between the inconsistent and the consistent proposer for each single experiment.

[11] Due to a programming error, the second distractor proposer offered four times \$1 and six times \$10 instead of five times \$1 and five times \$10 in Experiment 4a (expected value: \$6.4).

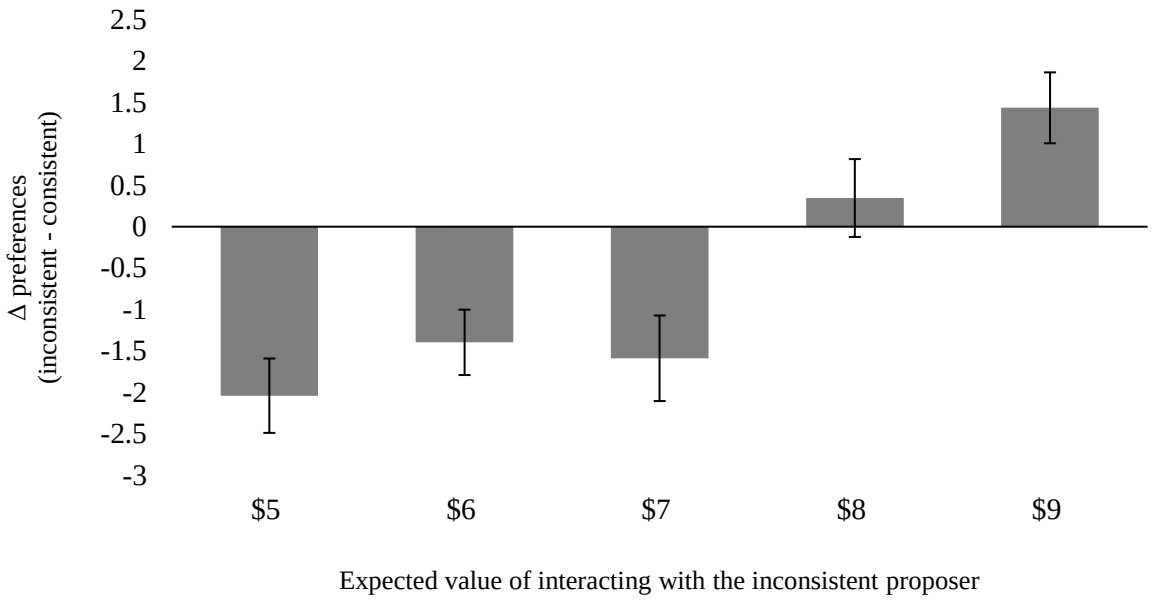

Figure 5. Stated proposer preferences in Experiments 4a–e. Bars depict the differences between the preference ratings for the inconsistent and the consistent proposers. Positive values indicate a preference for the inconsistent proposer. Negative values indicate a preference for the consistent proposer. Error bars represent *SEM*s.

3.6.3 Discussion

In a series of five single experiments, we pitted consistent against inconsistent offer behavior in an Ultimatum Game and systematically increased the expected value of bargaining with an inconsistent proposer. A preference for the consistent proposer emerged for equal expected values and when the expected value of interacting with the inconsistent proposer was increased to $6 and $7. Whereas stated preferences indicated statistical indifference when the expected value of interacting with the inconsistent proposer amounted to $8, the inconsistent proposer was significantly preferred with an expected value of $9. These findings imply an inconsistency premium of around 60%.

However, because stated preferences are not necessarily translatable into real decisions (see Cummings et al., 1995; Johannesson et al., 1998; Lambooij et al., 2015), we decided to conduct a fifth experiment that additionally assessed actual behavioral interaction choices.

3.7. Experiment 5

To corroborate our findings, we replicated Experiment 4 with an extended paradigm that implemented an incentivized choice between the consistent versus inconsistent proposer. In contrast to Experiment 4, Experiment 5 was realized as a full between-subjects design, with participants being randomly assigned to one of the five between-conditions. We thus had a 2 (Proposer behavior: consistent vs. inconsistent; within-subjects) X 5 (Expected value of interacting with the inconsistent proposer: $5 vs. $6 vs. $7 vs. $8 vs. $9; between-subjects) design.[12] The preregistration of Experiment 5 can be assessed at https://osf.io/9t32w/?view_only=3d1adc8aaadb4b9a9a4413c2b2d7ad8d.

3.7.1 Method

3.7.1.1 Participants. N = 250 participants were recruited via Amazon's *Mechanical Turk* and received $0.5 for compensation. We excluded data from participants who had already participated in one of the Experiments 4a–e (n = 11). Thus, the final sample size was N = 239 (96 female, 142 male, 1 gender-diverse; M_{age} = 35, SD_{age} = 11). All participants indicated their informed consent.

3.7.1.2 Procedure. The first part of Experiment 5 followed the same procedures as Experiment 4. However, after the preference ratings, participants chose a proposer (consistent vs. inconsistent; identified by the letter) to play one additional bargaining round. To put a real

[12] Since the two distractor proposers are conceptually irrelevant, we did not consider them in the experimental design. Including them for exploratory reasons in a 4 (Proposer behavior: consistent vs. inconsistent vs. distractor proposer 1 vs. distractor proposer 2; within-subjects) X 5 (Expected value of interacting with the inconsistent proposer: $5 vs. $6 vs. $7 vs. $8 vs. $9; between-subjects) analysis did, as expected, not yield substantial differences in the results. We therefore refrained from taking the distractor proposers into account in all further considerations.

price on predictability, this additional round of the game was incentivized. Having indicated their choices, participants were randomly presented one trial that corresponded to the offer strategy of the chosen proposer. As in the previous rounds, participants could decide whether to accept or reject the offer. At the end of the data collection, a random generator determined one participant who received the payoff from the extra round ($0 - $10) in addition to the regular compensation.

3.7.2 Results

3.7.2.1 Preference ratings. An ANOVA neither revealed a significant main effect of proposer behavior, $F(1,234) = 0.11$, $p = .75$, nor a significant main effect of the expected value, $F(4,234) = 1.08$, $p = .37$. Crucially, however, we found a significant interaction between the two factors, $F(4,234) = 3.63$, $p = .007$, $\eta_p^2 = .06$. Pairwise comparisons between the preference ratings for the consistent vs. inconsistent proposer for each of the between-conditions are depicted in Table 13 (see Appendix D). With equal expected values of $5, participants significantly preferred the consistent proposer to the inconsistent one. When the expected value of interacting with the inconsistent proposer was raised to $6 or $7, rating differences were not significant. A significant preference for the inconsistent proposer emerged only when their expected value was increased to $8. Surprisingly, we did not find significant differences between the preference ratings for the consistent vs. inconsistent proposer when the expected value of interacting with the inconsistent proposer amounted to $9. Figure 6 shows the overall preference pattern in terms of rating differences between the inconsistent and the consistent proposer for each between-condition.

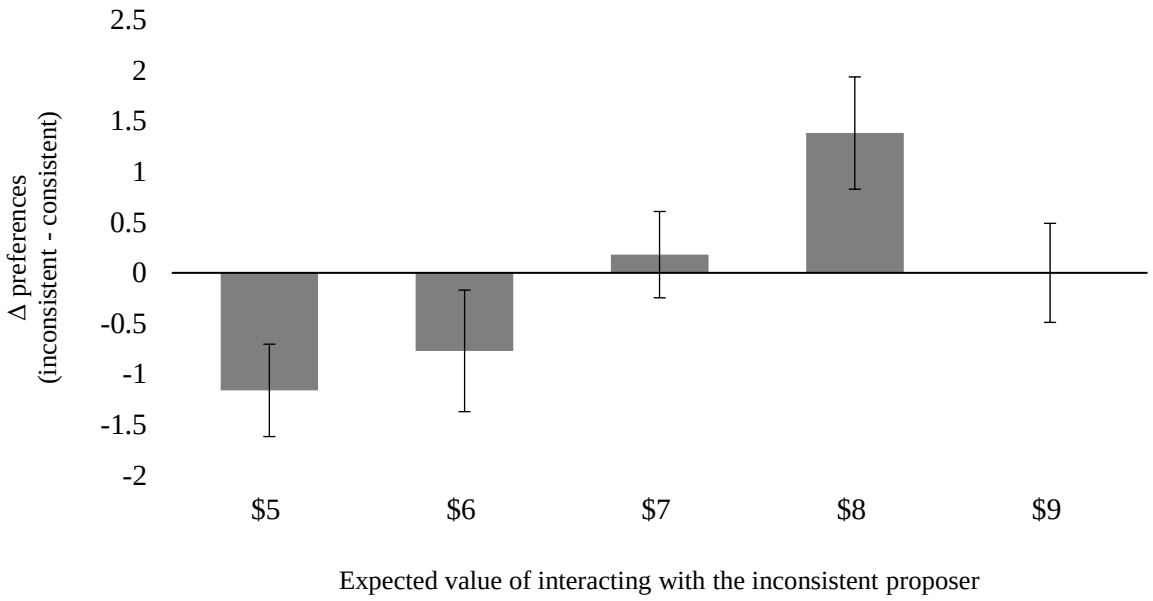

Figure 6. Stated proposer preferences in Experiment 5. Bars depict the differences between the preference ratings for the inconsistent and the consistent proposers. Positive values indicate a preference for the inconsistent proposer. Negative values indicate a preference for the consistent proposer. Error bars represent *SEM*s.

3.7.2.2 Proposer choice. A chi-square test for independence revealed a significant association between the expected value of interacting with the inconsistent proposer and proposer choice, $\chi^2(4) = 11.80$, $p = .019$, $V = .22$. However, separate binomial tests for each between-condition indicated that proposer choice proportions did not differ significantly from chance level in any condition (see Table 13, Appendix D).

3.7.3 Discussion

In our fifth experiment we observed a stated preference for the consistent proposer for equal expected values which only flipped after increasing the expected value of the interaction to $8. The indifference between the two proposers for the $7 condition implies an inconsistency premium of around 40%.

The indifference for the $9 condition might be a random spurious effect since it did not appear in the previous experiments. Yet, another possible explanation refers to the studied phenomenon of social interactions itself: Whereas the classical financial risk premium is derived from the trade-off between risk and return in a static and inanimate setting, our endeavor to investigate social settings naturally entails more complexity by introducing additional interpersonal dynamics, such as fairness norms or suspicion (e.g., Fehr & Schmidt, 1999; Steinel, van Beest, & van Dijk, 2013). Thus, despite having a substantially higher expected value, the inconsistent proposer might have been even less preferred because of participants' inequity aversion or suspicion about proposers behaving "too good to be true". Future research may investigate this question in more detail.

With a view to revealed preferences, choice proportions did not significantly differ from chance in any condition. Thus, although the overall association between the expected value of interacting with the inconsistent proposer and proposer choice was significant, our results suggest indifference between the two proposers when directly comparing choices. A reason for this pattern may be that participants could always reject a $0 offer from the inconsistent proposer. Consequently, the expected value might not capture the immaterial reward of punishing unfairness which would manifest in a bias favoring inconsistency. We will come back to discussing this point in the General Discussion.

To obtain clearer results we decided to streamline our paradigm by omitting the two distractor players, thus focusing on the crucial comparison between the consistent and the inconsistent proposer.

3.8 Experiment 6

Eliminating potential distractor interferences by putting the consistent and the inconsistent proposer into direct competition allows for a clearer view on participants' actual preferences. Therefore, Experiment 6 was a pre-registered replication of Experiment 5 with the

sole modification that we did not implement the two distractor proposers. The preregistration can be assessed at https://osf.io/9uq3s/?view_only=f7b45f2be3634863b69f6395c646f598.

3.8.1 Method

3.8.1.1 Participants. N = 251 participants were recruited via Amazon's *Mechanical Turk* and received \$0.3 for compensation. We excluded data from participants who did not pass the attention check (identifying the color of a word; n = 9). Thus, the final sample size was N = 242 (96 female, 145 male, 1 gender-diverse; M_{age} = 33, SD_{age} = 10).

3.8.1.2 Procedure. The procedure was similar to the previous experiment, with the crucial difference of presenting only offers from the consistent and the inconsistent proposer. This reduced the sequence of Ultimatum Games to 20 rounds in total. Having completed these rounds, participants were asked to indicate how much they would like to bargain further rounds with each of the two proposers on a scale from 0 (*not at all*) to 10 (*very much*). The incentivized choice procedure was the same as in Experiment 5.

3.8.2 Results

3.8.2.1 Preference ratings. We did neither find a significant main effect of proposer behavior, $F(1,237) = 2.65$, $p = .11$, nor of the expected value of interacting with the inconsistent proposer, $F(4,237) = 1.27$, $p = .28$, but, as expected, a significant interaction between the two factors, $F(4,237) = 8.15$, $p < .001$, $\eta_p^2 = .12$. Pairwise comparisons between the preference ratings for the consistent vs. inconsistent proposer for each of the between-conditions are depicted in Table 14 (see Appendix D). Participants significantly preferred the consistent proposer to the inconsistent one with equal expected values of \$5 and when the expected value of interacting with the inconsistent proposer was \$6. When the expected value of the interaction amounted to \$7, \$8, or \$9, the preference was reversed. Figure 7 plots the overall preference

pattern in terms of rating differences between the inconsistent and the consistent proposer for each between-condition.

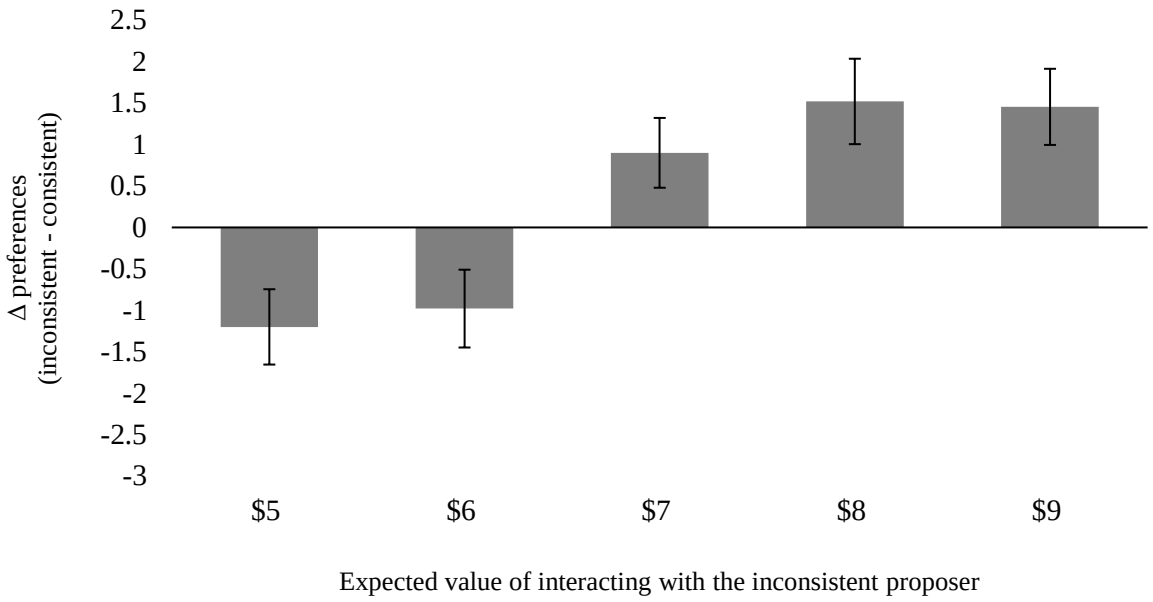

Figure 7. Stated proposer preferences in Experiment 6. Bars depict the differences between the preference ratings for the inconsistent and the consistent proposers. Positive values indicate a preference for the inconsistent proposer. Negative values indicate a preference for the consistent proposer. Error bars represent *SEM*s.

3.8.2.2 Proposer choice. A chi-square test for independence indicated a significant association between the expected value of interacting with the inconsistent proposer and proposer choice, $\chi^2(4) = 14.53$, $p = .006$, $V = .25$. We did not observe significant differences from chance level for equal expected values and when the expected value of interacting with the inconsistent proposer was \$6 or \$7 (see Appendix D, Table 14). However, proposers significantly preferred the inconsistent proposer when the expected value of the interaction was increased to \$8 or \$9.

3.8.3 Discussion

Streamlining our experimental design by putting the consistent and the inconsistent proposer into direct competition revealed significant stated preferences for the consistent proposer until the expected value of interacting with the inconsistent proposer was increased to $7.

Omitting the two distractor players from our paradigm also produced clearer results regarding participants' revealed preferences as measured by proposer choice. Whereas the results obtained in Experiment 5 suggest indifference across all expected value conditions, the present findings indicate a significant preference for the inconsistent proposer once the expected value amounted to $8 or $9. That is, opposed to the assumption of inconsistency aversion, we again did not observe a significant revealed preference for the consistent proposer even when the expected values of both proposers were equal. Hence, also in a streamlined setting, participants were not per se willing to forgo part of their potential returns to interact with a consistent proposer when it came to concrete behavioral choices. To a certain degree, this could imply that the immaterial reward of punishing unfairness might have triggered a bias favoring inconsistency. However, Experiment 6 provides first evidence that incentivized preferences shift in favor of an inconsistent proposer once the expected value of the interaction noticeably exceeds the expected value of interacting with the consistent one. Overall, we locate the inconsistency premium at around 20%.

Importantly, Experiments 1–6 were designed in such a way that an inconsistent partner's behavioral inconsistency decreased with increasing expected values (see Appendix D, Tables 9–14). Therefore, so far, we cannot clearly determine to what extent our current inconsistency premium truly estimates the required difference in expected values, and to what extent it is skewed by the additional variance reduction. Experiment 7 removed this confound.

3.9 Experiment 7

The results obtained in the previous experiments yield a first estimate for the inconsistency premium of about 33%. However, expected value and variance were not manipulated independently from each other in these experiments. Rather, variance decreased with increasing expected values. Given this confound, we got a biased estimate for the tipping point of preferences. Experiment 7 fully disentangled expected return and behavioral inconsistency to test whether we actually underestimated the premium required to make decision-makers indifferent towards inconsistency.

To achieve a more fine-grained resolution of the inconsistency premium, we increased the amount of money that has to be divided and endowed the proposers with $20. Whereas the consistent proposer always offered half of the amount in each trial (yielding an expected value of $10), we manipulated the expected value of interacting with the inconsistent proposer across six between-participants conditions while keeping the variances constant (see Table 15, Appendix D). As before, the inconsistent proposers made offers that were higher than the consistent proposers' offers and offers that were lower. Crucially, expected values were manipulated by increasing both the low and high offers of the inconsistent proposers by $1 across the different between-conditions. That is, constant variances were ensured by keeping the difference between low and high offers constant at $10 (see Table 15, Appendix D). The preregistration of Experiment 7 can be assessed at https://osf.io/qcrx6/?view_only=99fe9c5c6b414c749debeb9abc6ff3bd.

3.9.1 Method

3.9.1.1 Participants. N = 781 participants (384 female, 391 male, 6 gender-diverse; M_{age} = 33, SD_{age} = 10) were recruited via Amazon's *Mechanical Turk* and received $0.5 for compensation.

3.9.1.2 Procedure. Experiment 7 followed the same procedure as Experiments 4–6. However, whereas the consistent proposer still offered $10 in each trial, the inconsistent proposer made a low offer in four trials and a high offer in six trials (see Table 15, Appendix D). Stated and revealed preference assessments were the same as in Experiments 4–6.

3.9.2 Results

3.9.2.1 Preference ratings. Consistent with the previous experiments, an ANOVA revealed neither a significant main effect of proposer behavior, $F(1,775) = 0.12$, $p = .73$, nor of the expected value of interacting with the inconsistent proposer, $F(5,775) = 1.73$, $p = .13$. Most importantly, however, we found a significant interaction between both factors, $F(5,775) = 8.89$, $p < .001$, $\eta_p^2 = .05$. Pairwise comparisons between the preference ratings for the consistent versus inconsistent proposer for each of the between-conditions are depicted in Table 15 (see Appendix D). Parallel to the previous findings, participants significantly preferred the consistent proposer when the expected value of interacting with the inconsistent proposer was $10 to $12. However, once the expected value of the interaction with the inconsistent proposer amounted to at least $14, preferences flipped in favor of the inconsistent proposer. Figure 8 depicts the overall preference pattern in terms of rating differences between the inconsistent and the consistent proposer for each between-condition.

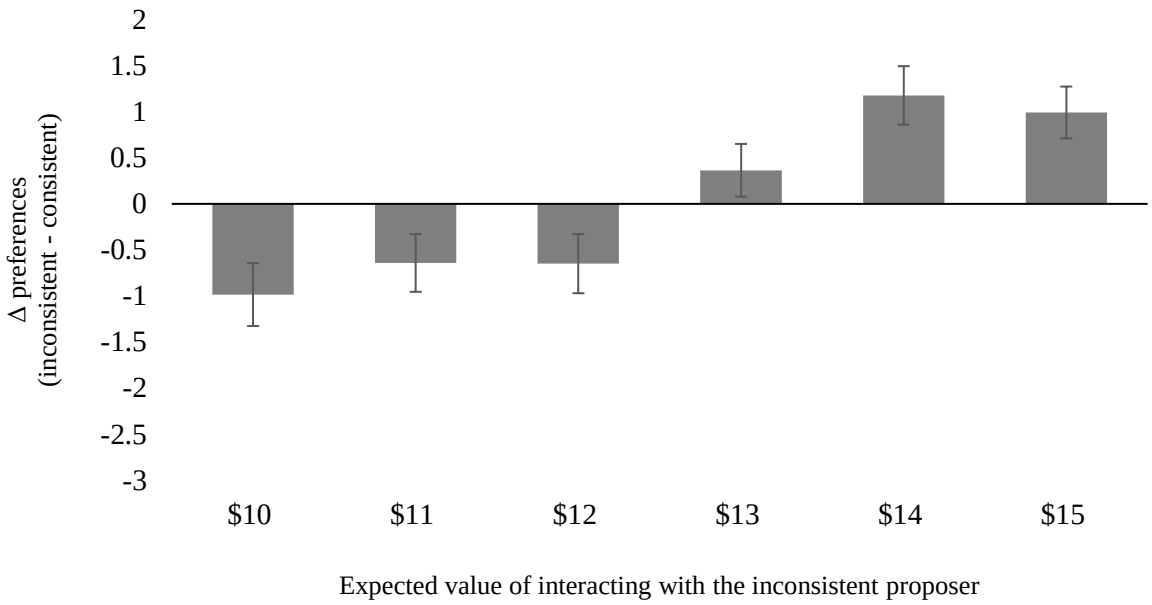

Figure 8. Stated proposer preferences in Experiment 7. Bars depict the differences between the preference ratings for the inconsistent and the consistent proposers. Positive values indicate a preference for the inconsistent proposer. Negative values indicate a preference for the consistent proposer. Error bars represent *SEM*s.

3.9.2.2 Proposer choice. A chi-square test for independence indicated a significant association between the expected value of interacting with the inconsistent proposer and proposer choice, $\chi^2(5) = 30.26$, $p < .001$, $V = 0.20$. In detail, there was a significant consistency preference for equal expected values of \$10 (see Table 15, Appendix D). Participants were indifferent when the expected value of interacting with the inconsistent proposer was increased to \$11, \$12, or \$13. Yet, in contrast, participants significantly preferred the inconsistent proposer once the expected value of the interaction amounted to \$14 or \$15.

3.9.3 Discussion

Experiment 7 addressed a shortcoming of the previous experiments: the confound between risk and return. Whereas in Experiments 1–6, interacting with an inconsistent partner became more and more profitable across conditions but also concurrently less risky, our novel experiment kept the level of riskiness constant and manipulated only the expected value of the inconsistent interaction. In the adjusted setting, we locate the inconsistency premium at around 20%, which is broadly in line with our previous results.

3.10 General Discussion

The history of economic thought suggests that decision-makers trade off the potential risks and returns of an investment option. Furthermore, decision-makers are generally assumed to prefer lower risks to higher ones and high returns to lower ones. Bridging the literatures of economics and social psychology, the present research is the first to transfer this trade-off principle to the social-psychological domain by investigating human decision-making in social interactions under strategic uncertainty. Our aim was to estimate the inconsistency premium required to make decision-makers prefer a more profitable, but inconsistent partner to a consistent, but less profitable one. This allows us to quantify the psychological value of consistency and the payoff that people are willing to forgo to avoid uncertainty and to ensure consistency in social interactions.

In a set of seven experiments, we assessed the magnitude of this inconsistency premium by pitting consistent against inconsistent behaviors while varying the expected returns in both an organizational workplace setting (Experiments 1–3) and in an economic game (Experiments 4–7). Also, we measured both stated interaction preferences (Experiments 1, 4–7), explicit ratings of co-worker liking (Experiment 3) and revealed interaction preferences (Experiments 2, 5–7).

In a nutshell, our findings suggest that people indeed exhibit a psychological preference for consistency and are willing to forgo substantial benefits to avoid inconsistency. Investigating the risk-return trade-off in a workplace setting in Experiments 1 and 2 suggests an inconsistency premium estimate of around 32% of the total benefits from interacting with a consistent co-worker. Experiment 3 examined the effects of expected value versus variance on more general social evaluations, providing first evidence that also ratings of co-worker liking are determined by an interplay of both these factors. This affective inconsistency premium amounts to around 16%. Experiments 4–7 sought to generalize our findings to a more abstract and formalized context. Therefore, we operationalized risk and return by offer consistency and potential gains in a repeated Ultimatum Game. This economic bargaining setting reveals an inconsistency premium estimate of about 35%. Collapsed across all present experiments, settings, and scales, we thus locate the inconsistency premium at around 31%. This implies that participants were overall willing to waive up to 24% of their potential returns from interacting with a more profitable but inconsistent person to "buy" consistency in social interactions.

To a certain degree, if inconsistency (i.e., variance) decreases with increasing returns, the inconsistency premium should be underestimated. However, the estimates from Experiments 1–6, where expected returns were confounded with behavioral consistency, did not substantially deviate from the estimate obtained in the unconfounded setting of Experiment 7. Therefore, the magnitude of behavioral inconsistencies does not seem to affect preferences. Rather, the current findings give reason to assume that consistency is psychologically perceived in a dichotomous way – with someone either behaving consistently or not –, while the exact inconsistency magnitude is less relevant. It remains an open question whether people are not capable of responding to different gradients of consistency, or whether they are not willing to do so and, as cognitive misers, rather rely on simple binary categorization heuristics (see also Kahneman, 2011; Simon, 1955; Strack & Deutsch, 2004). Reversing our current experimental setup by manipulating only the magnitude of the inconsistent partners' behavioral consistency

while keeping their expected values constant across conditions would present a promising first step towards investigating this question more closely – if people are indeed insensitive to the consistency magnitude, we would expect no differences in preferences between the different conditions. Future research may thus address this issue in more detail.

In line with previous literature from both the economic and psychological domain, the present work focused mainly on the aversive attributes of inconsistency. What has been largely ignored is that inconsistency might in fact reveal a Janus face, with a lack of certainty, control, and plannability looming on the one side, but the promise of thrill, excitement, and diversification on the other one (see also Simandan, 2018; Sinaceur, Adam, van Kleef, & Galinsky, 2013). Supporting this duality, we found a significant main effect of the co-worker's behavioral consistency on basic social evaluations of person liking in Experiment 3, indicating that the inconsistent co-worker was on average liked *more* than the consistent one. Although qualified by a significant interaction, this main effect did not show in settings that assessed participants' behavior or intentions, thus pointing towards different mechanisms underlying these different evaluations. At this point, we can only speculate about the psychological mechanisms underpinning these patterns. Yet, our current findings suggest an evaluation-interaction discrepancy of peoples' social (in)consistency preferences (for a similar phenomenon, see the *choice-judgment discrepancy*, Tversky & Griffin, 1991; see also Montgomery, Selart, Gärling, & Lindberg, 1994; Tassy, Oullier, Mancini, & Wicker, 2013). Per se, inconsistency seems to be affectively interesting – unless it entails manifest consequences that affect one's own stake. In contrast, when it comes to social interactions that serve as means to a personal end (whether this is in terms of increasing work-related benefits or financial payoff), people rather prefer behavioral consistency to ensure successful goal achievement.

Even though stated and revealed preference patterns coincided for social interactions in a workplace setting, there was no reliable consistency preference when actually choosing a partner in the Ultimatum Game. To a considerable degree, these results might reflect the idiosyncrasies of the Ultimatum Game because participants can always punish proposers by rejecting their offer. Previous research suggests that punishing has a motivational value on its own since it activates neural pathways that are associated with reward processing (de Quervain et al., 2004; see also Fehr & Gächter, 2002; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). In Experiments 5 and 6, the anticipated pleasure of potentially punishing an inconsistent proposer might have counteracted the anticipated pleasure of a certain return promised by choosing the consistent proposer. Nonetheless, participants' stated preferences clearly reveal that people could be willing to forgo a substantial part of their expected returns for interacting with a consistent (compared to an inconsistent) partner in an economic game.

By and large, the present findings lend support to the notion that the psychological equivalents of variance and expected value antagonistically impact social evaluations and decision-making under strategic uncertainty. The trade-off between the benefits of an interaction and the behavioral inconsistency of an interaction partner has hitherto been scarcely considered in psychological models, and our results firstly provide an estimate for the price that people are willing to pay for predictability. Our current estimate of 31% is, of course, limited to the presently realized experimental settings of workplace collaborations and interactions in an economic game, and thus, we see wide scope for future research that investigates the context sensitivity and generalizability of this premium to further interaction settings (e.g., romantic relationships) or interaction durations (e.g., ten instead of one further round), different return instantiations (e.g., time) and more realistic contexts (e.g., face-to-face interactions in the lab). Furthermore, although Amazon's *Mechanical Turk* provides a very heterogeneous and diverse participant population and to a certain degree already allows the generalization of our results, further studies may aim to substantiate our estimate by replicating the present experiments with

non-US samples. In a related vein, previous research suggests that even though risk preferences are considerably heterogeneous across countries (e.g., Rieger, Wang, & Hens, 2014), these differences primarily relate to different perceptions of risks rather than differences in attitudes towards the perceived risk (Weber & Hsee, 1998). To a considerable degree, the behavioral inconsistency in our experiments was less open to subjective perceptions but rather objective such that we manipulated the actual variance in observable behavior. Therefore, we expect that the size of the premium should be rather similar across different samples. While future research might shed further light on this issue, we are certain that our findings present a valuable first step towards gauging the price that people are willing to pay for a consistent interaction partner.

In conclusion, the current research points out a psychological preference for consistency, with people being willing to forgo a substantial part of their potential returns to avoid uncertainty and to ensure predictability in social interactions. The present research thereby contributes to harvesting the synergies of economic and psychological approaches to social interactions and decision-making, creating bridges and fostering an interdisciplinary scientific understanding of human behavior.

3.11 Appendix

3.11.1 Appendix D

Table 9

Experiment 1: Manipulations, samples and results.

Sample	Inconsistent co-worker: Expected value (EV) and variance (Var)	Statistics for the pairwise comparison between the ratings for the consistent vs. inconsistent co-worker
n = 51	EV: 50% (5 x 10%; 5 x 90%) Var: 1777.78	M_{cons} = 7.04, SE_{cons} = 0.23 M_{incons} = 4.98, SE_{incons} = 0.33 $t(50) = 4.52$, $p < .001$, $d_z = 0.63$, 95% CI = [1.14; 2.97]
n = 51	EV: 58% (4 x 10%; 6 x 90%) Var: 1706.67	M_{cons} = 6.12, SE_{cons} = 0.37 M_{incons} = 5.47, SE_{incons} = 0.38 $t(50) = 0.95$, $p = .35$, $d_z = 0.13$, 95% CI = [-0.72, 2.02]
n = 52	EV: 66% (3 x 10%; 7 x 90%) Var: 1493.33	M_{cons} = 6.56, SE_{cons} = 0.32 M_{incons} = 6.37, SE_{incons} = 0.33 $t(51) = 0.38$, $p = .71$, $d_z = 0.05$, 95% CI = [-0.82, 1.21]
n = 48	EV: 74% (2 x 10%; 8 x 90%) Var: 1137.78	M_{cons} = 5.58, SE_{cons} = 0.40 M_{incons} = 6.71, SE_{incons} = 0.41 $t(47) = -1.60$, $p = .12$, $d_z = -0.23$, 95% CI = [-2.54, 0.29]
n = 48	EV: 82% (1 x 10%; 9 x 90%) Var: 640	M_{cons} = 4.65, SE_{cons} = 0.31 M_{incons} = 7.94, SE_{incons} = 0.26 $t(47) = -6.82$, $p < .001$, $d_z = -0.98$, 95% CI = [-4.26, -2.32]

Table 10

Experiment 2: Manipulations, samples and results.

Sample	Inconsistent co-worker: Expected value (EV) and variance (Var)	Choice proportions for the consistent vs. inconsistent co-worker
n = 52	EV: 50% (5 x 10%; 5 x 90%) Var: 1777.78	consistent 75.0% (39) inconsistent 25.0% (13) *p* < .001
n = 52	EV: 58% (4 x 10%; 6 x 90%) Var: 1706.67	consistent 55.8% (29) inconsistent 44.2% (23) *p* = .49
n = 51	EV: 66% (3 x 10%; 7 x 90%) Var: 1493.33	consistent 45.1% (19) inconsistent 54.9% (31) *p* = .58
n = 49	EV: 74% (2 x 10%; 8 x 90%) Var: 1137.78	consistent 42.9% (21) inconsistent 57.1% (28) *p* = .39
n = 47	EV: 82% (1 x 10%; 9 x 90%) Var: 640	consistent 19.1% (9) inconsistent 80.9% (38) *p* < .001

Table 11

Experiment 3: Manipulations, samples and results.

Sample	Inconsistent co-worker: Expected value (EV) and variance (Var)	Statistics for the pairwise comparison between the ratings for the consistent vs. inconsistent co-worker
n = 50	EV: 50% (5 x 10%; 5 x 90%) Var: 1777.78	M_{cons} = 6.64, SE_{cons} = 0.25 M_{incons} = 5.84, SE_{incons} = 0.31 $t(49) = 1.88$, $p = .07$, $d_z = 0.27$, 95% CI = [-0.06, 1.66]
n = 51	EV: 58% (4 x 10%; 6 x 90%) Var: 1706.67	M_{cons} = 6.39, SE_{cons} = 0.26 M_{incons} = 6.28, SE_{incons} = 0.27 $t(50) = 0.34$, $p = .73$, $d_z = 0.05$, 95% CI = [-0.58, 0.81]
n = 50	EV: 66% (3 x 10%; 7 x 90%) Var: 1493.33	M_{cons} = 5.92, SE_{cons} = 0.21 M_{incons} = 6.66, SE_{incons} = 0.26 $t(49) = -2.38$, $p = .02$, $d_z = -0.34$, 95% CI = [-1.37, -0.11]
n = 48	EV: 74% (2 x 10%; 8 x 90%) Var: 1137.78	M_{cons} = 6.25, SE_{cons} = 0.26 M_{incons} = 6.98, SE_{incons} = 0.28 $t(47) = -1.78$, $p = .08$, $d_z = -0.26$, 95% CI = [-1.55, 0.09]
n = 52	EV: 82% (1 x 10%; 9 x 90%) Var: 640	M_{cons} = 5.35, SE_{cons} = 0.25 M_{incons} = 8.10, SE_{incons} = 0.19 $t(51) = -8.53$, $p < .001$, $d_z = -1.18$, 95% CI = [-3.40, -2.10]

Table 12

Experiments 4a–e: Sample, manipulations and results.

Experiment	Sample	Inconsistent proposer: Expected value (EV) and variance (Var)	Statistics for the pairwise comparison between the ratings for the consistent vs. inconsistent proposer
4a	N = 50 21 women, 28 men, 1 gender-diverse M_{age} = 35 years, SD_{age} = 10	EV: \$5 (5 x \$0; 5 x \$10) Var: 27.78	M_{cons} = 6.88, SE_{cons} = 0.32 M_{incons} = 4.84, SE_{incons} = 0.42 $t(49) = 4.54$, $p < .001$, $d_z = 0.64$, 95% CI [1.14, 2.94]
4b	N = 48 17 women, 31 men M_{age} = 36 years, SD_{age} = 9	EV: \$6 (4 x \$0; 6 x \$10) Var: 26.67	M_{cons} = 6.15, SE_{cons} = 0.31 M_{incons} = 4.75, SE_{incons} = 0.36 $t(47) = 3.54$, $p = .001$, $d_z = 0.51$, 95% CI [0.60, 2.19]
4c	N = 51 25 women, 25 men, 1 gender-diverse M_{age} = 34 years, SD_{age} = 10	EV: \$7 (3 x \$0; 7 x \$10) Var: 23.33	M_{cons} = 7.12, SE_{cons} = 0.29 M_{incons} = 5.53, SE_{incons} = 0.36 $t(50) = 3.08$, $p = .003$, $d_z = 0.48$, 95% CI [0.55, 2.63]
4d	N = 52 15 women, 37 men M_{age} = 33 years, SD_{age} = 8	EV: \$8 (2 x \$0; 8 x \$10) Var: 17.78	M_{cons} = 6.58, SE_{cons} = 0.37 M_{incons} = 6.92, SE_{incons} = 0.37 $t(51) = -0.74$, $p = .47$, $d_z = -0.10$, 95% CI [-1.29, 0.60]
4e	N = 46 19 women, 27 men M_{age} = 34 years, SD_{age} = 9	EV: \$9 (1 x \$0; 9 x \$10) Var: 10	M_{cons} = 5.72, SE_{cons} = 0.45 M_{incons} = 7.15, SE_{incons} = 0.43 $t(45) = -3.35$, $p = .002$, $d_z = -0.49$, 95% CI [-2.30, -0.57]

Table 13

Experiment 5: Sample, manipulations and results.

Sample	Inconsistent proposer: Expected value (EV) and variance (Var)	Statistics for the pairwise comparison between the ratings for the consistent vs. inconsistent proposer	Choice proportions for the consistent vs. inconsistent proposer
n = 45	EV: \$5 (5 x \$0; 5 x \$10) Var: 27.78	M_{cons} = 6.47, SE_{cons} = 0.42 M_{incons} = 5.31, SE_{incons} = 0.46 $t(44) = 2.53$, $p = .015$, $d_z = 0.38$, 95% CI [0.24, 2.08]	consistent 62.2% (28) inconsistent 37.8% (17) $p = .14$
n = 48	EV: \$6 (4 x \$0; 6 x \$10) Var: 26.67	M_{cons} = 6.02, SE_{cons} = 0.44 M_{incons} = 5.25, SE_{incons} = 0.40 $t(47) = 1.28$, $p = .21$, $d_z = 0.19$, 95% CI [-0.44, 1.98]	consistent 62.5% (30) inconsistent 37.5% (18) $p = .11$
n = 50	EV: \$7 (3 x \$0; 7 x \$10) Var: 23.33	M_{cons} = 5.86, SE_{cons} = 0.38 M_{incons} = 6.04, SE_{incons} = 0.39 $t(49) = -0.42$, $p = .68$, $d_z = -0.06$, 95% CI [-1.04, 0.68]	consistent 38.0% (19) inconsistent 62.0% (31) $p = .12$
n = 48	EV: \$8 (2 x \$0; 8 x \$10) Var: 17.78	M_{cons} = 5.46, SE_{cons} = 0.44 M_{incons} = 6.83, SE_{incons} = 0.42 $t(47) = -2.48$, $p = .017$, $d_z = -0.36$, 95% CI [-2.49, -0.26]	consistent 37.5% (18) inconsistent 62.5% (30) $p = .11$
n = 48	EV: \$9 (1 x \$0; 9 x \$10) Var: 10	M_{cons} = 6.52, SE_{cons} = 0.35 M_{incons} = 6.52, SE_{incons} = 0.37 $t(47) < .001$, $p > .99$, $d_z < .001$, 95% CI [-0.98, 0.98]	consistent 45.8% (22) inconsistent 54.2% (26) $p = .67$

Table 14

Experiment 6: Sample, manipulations and results.

Sample	Inconsistent proposer: Expected value (EV) and variance (Var)	Statistics for the pairwise comparison between the ratings for the consistent vs. inconsistent proposer	Choice proportions for the consistent vs. inconsistent proposer
n = 50	EV: \$5 (5 x \$0; 5 x \$10) Var: 27.78	M_{cons} = 7.10, SE_{cons} = 0.32 M_{incons} = 5.90, SE_{incons} = 0.39 $t(49) = 2.64$, $p = .011$, $d_z = 0.37$, 95% CI [0.29, 2.11]	consistent 58.0% (29) inconsistent 42.0% (21) $p = .32$
n = 48	EV: \$6 (4 x \$0; 6 x \$10) Var: 26.67	M_{cons} = 7.40, SE_{cons} = 0.34 M_{incons} = 6.42, SE_{incons} = 0.41 $t(47) = 2.09$, $p = .042$, $d_z = 0.30$, 95% CI [0.04, 1.92]	consistent 58.3% (28) inconsistent 41.7% (20) $p = .31$
n = 50	EV: \$7 (3 x \$0; 7 x \$10) Var: 23.33	M_{cons} = 6.26, SE_{cons} = 0.41 M_{incons} = 7.16, SE_{incons} = 0.40 $t(49) = -2.14$, $p = .038$, $d_z = -0.30$, 95% CI [-1.75, -0.05]	consistent 44.0% (22) inconsistent 56.0% (28) $p = .48$
n = 50	EV: \$8 (2 x \$0; 8 x \$10) Var: 17.78	M_{cons} = 6.18, SE_{cons} = 0.43 M_{incons} = 7.70, SE_{incons} = 0.40 $t(49) = -2.95$, $p = .005$, $d_z = -0.42$, 95% CI [-2.56, -0.49]	consistent 30.0% (15) inconsistent 70.0% (35) $p = .007$
n = 44	EV: \$9 (1 x \$0; 9 x \$10) Var: 10	M_{cons} = 6.71, SE_{cons} = 0.38 M_{incons} = 8.16, SE_{incons} = 0.29 $t(43) = -3.17$, $p = .003$, $d_z = -0.48$, 95% CI [-2.38, -0.53]	consistent 31.8% (14) inconsistent 68.2% (30) $p = .023$

Table 15

Experiment 7: Sample, manipulations and results.

Sample	Inconsistent proposer: Expected value (EV) and variance (Var)	Statistics for the pairwise comparison between the ratings for the consistent vs. inconsistent proposer	Choice proportions for the consistent vs. inconsistent proposer
n = 130	EV: \$10 (4 x \$4; 6 x \$14) Var: 26.67	M_{cons} = 6.45, SE_{cons} = 0.26 M_{incons} = 5.47, SE_{incons} = 0.25 $t(129) = 2.89$, $p = .005$, $d_z = 0.25$, 95% CI [0.31, 1.66]	consistent 60.8% (79) inconsistent 39.2% (51) $p = .018$
n = 131	EV: \$11 (4 x \$5; 6 x \$15) Var: 26.67	M_{cons} = 6.69, SE_{cons} = 0.23 M_{incons} = 6.05, SE_{incons} = 0.23 $t(130) = 2.05$, $p = .04$, $d_z = 0.18$, 95% CI [0.02, 1.26]	consistent 48.1% (63) inconsistent 51.9% (68) $p = .73$
n = 131	EV: \$12 (4 x \$6; 6 x \$16) Var: 26.67	M_{cons} = 6.66, SE_{cons} = 0.24 M_{incons} = 6.01, SE_{incons} = 0.24 $t(130) = 2.02$, $p = .05$, $d_z = 0.18$, 95% CI [0.01, 1.28]	consistent 53.4% (70) inconsistent 46.6% (61) $p = .49$
n = 129	EV: \$13 (4 x \$7; 6 x \$17) Var: 26.67	M_{cons} = 6.39, SE_{cons} = 0.25 M_{incons} = 6.75, SE_{incons} = 0.22 $t(128) = -1.27$, $p = .21$, $d_z = -0.11$, 95% CI [-0.93, 0.20]	consistent 44.2% (57) inconsistent 55.8% (72) $p = .22$
n = 130	EV: \$14 (4 x \$8; 6 x \$18) Var: 26.67	M_{cons} = 5.85, SE_{cons} = 0.27 M_{incons} = 7.03, SE_{incons} = 0.23 $t(129) = -3.71$, $p < .001$, $d_z = -0.33$, 95% CI [-1.81, -0.55]	consistent 33.1% (43) inconsistent 66.9% (87) $p < .001$
n = 130	EV: \$15 (4 x \$9; 6 x \$19) Var: 26.67	M_{cons} = 6.16, SE_{cons} = 0.24 M_{incons} = 7.15, SE_{incons} = 0.23 $t(43) = -3.53$, $p = .001$, $d_z = -0.31$, 95% CI [-1.54, -0.44]	consistent 34.6% (45) inconsistent 65.4% (85) $p = .001$

Chapter 4 – Discussion

Under the umbrella of unpredictability, the research presented in the current dissertation split into two streams: the unpredicted and the unpredictable. In the following sections, I will outline in how far the evidence obtained contributes to clarifying this dissertation's three central questions on the causes of surprise, the graded-ness of multiple surprise responses, and the valence of the unpredicted and the unpredictable. The discussion of overall implications and limitations will finally terminate with a conclusion. Note that although the research presented offers a lot of material to further elaborate on, the following paragraphs will focus on discussing the main questions defined in the introduction for the sake of stringency.

In Chapter 2, I investigated the currently competing accounts on the causal mechanisms of surprise. Specifically, I developed and validated an experimental paradigm that assessed the effects of both unexpectedness (as operationalized by the deviance between an event and the previous mode of events) and the ease of sense-making (as operationalized by the strength of expectation constraints) on the behavioral, affective, experiential, and cognitive surprise responses. The results revealed significant effects of deviance and an only unsystematic influence of expectation constraints, pointing towards a dominant causal role of unexpectedness for surprise.

Furthermore, I pitched a dichotomous all-or-nothing account against the assumption of continuous grades of surprise to explore whether all surprising events trigger the same uniform response patterns. While varying the ease of making sense of an event neither revealed dichotomous nor graded effects, the evidence on the effects of the degree of deviance supports a dichotomous view: Response intensities differed between deviance and non-deviance of an event, but not between a medium and high degree of deviance.

Depending on the theory and evidence applied, responses to the unpredicted and the unpredictable resemble an affective "chameleon" (Noordewier & Breugelmans, 2013, p.1327),

at one time being conceptualized as positive, another time as negative, and yet another time as essentially valence-free. In the current dissertation, I took a differentiated view on the valence of both these phenomena. In Chapter 2, I investigated the valence of the unpredicted by assessing explicit liking ratings for surprising events. In Chapter 3, I derived the valence of the unpredictable from the psychological value that people attach to predictability in social interactions. My findings revealed unsystematic responses to surprising events, implying that the unpredicted is inherently valence-free. Moreover, I demonstrated a fundamental averseness of the unpredictable, which can, however, be reversed by the prospect of high returns of an interaction.

4.1 Limitations, Implications, and Future Directions

Offering important insights into the nature of the unpredicted and the unpredictable, the current findings refine our theoretical understanding of these two phenomena und may equally spark new lines of research. Nonetheless, they should be conceptualized only as a beginning, holding room for further improvement. In the following paragraphs, I will elaborate on the implications and limitations of the present research and point out future directions.

4.1.1 The Cause of Surprise: Unbundling the Confusion

In Chapter 2, I demonstrated that the behavioral, experiential, and cognitive surprise responses are influenced by the degree of deviance whereas the strength of expectation constraints had only unsystematic effects. If participants were in search of explanations for deviant events, they should have taken into account the a priori induced expectations, since these provide (depending on the condition) at least partial information to resolve the surprise. Since I did not observe a significant impact of expectation constraints, I concluded that the key driving principle of surprise is unexpectedness and not the ease of making sense of an event.

Likewise, however, these findings underscore the relevance and necessity of delimiting a clear definitional understanding of what constitutes an expectation and, in consequence, an unexpected event. While the evidence reported bolsters a repetition-change definition of "unexpected" (i.e., as "an unannounced deviation from the previous mode of presentation", Meyer et al., 1997, p. 257; see also Chapter 1.1), it does *not* support a (leading) role of a priori expectations. Surprise seems to be less about in-advance induced and fixed beliefs and more about implicitly forming "plastic" expectations that transform on a trial-by-trial basis over the course of events. This appears largely convincing from an evolutionary perspective as only such a flexible prediction redesign enables efficient adaptation in the long run (Itti & Baldi, 2009; see also Friston, Thornton, & Clark, 2012). However, unexpectedness approaches might have to reconsider the scope of what kind of expectations, beliefs, and schemata their theory can actually account for.

Although the presently depicted findings speak in favor of an unexpectedness account of surprise, I deem it important to go beyond this conclusion and outline the potential reconcilableness of both perspectives. Unexpectedness and sense-making approaches are not necessarily mutually exclusive and doomed to a rigid "either/or", but conceivably rather result from a confusion of different temporal perspectives and terminological interpretations of surprise: the ones focusing on the processes "on-line", the others placing emphasis on post-hoc mechanisms (Munnich et al., 2019; see also Loewenstein, 2019). Pezzo's (2003) theory on hindsight bias might offer an (unintended) solution approach by distinguishing between "initial surprise" as the immediate response to unexpected events, and "resultant surprise" as the conscious outcome of sense-making operations (see Munnich et al., 2019, for a more detailed explanation; see also Noordewier et al., 2016, for a temporal account on surprise). Hence, both accounts may pursue an equally appropriate approach to surprise – and might profit even more from being merged into a synergy-creating unified perspective.

4.1.2 The Structure of Surprise: Grading the Dichotomy

With a view to the structure of reactions to the unpredicted, the evidence obtained in Chapter 2 points towards dichotomous responses that differ between deviant and non-deviant events but are not sensitive to further deviance gradations. From an evolutionary perspective, the main advantage of such a response dichotomy may be its information efficiency. Assuming that the purpose of surprise is to facilitate the mastering of an event, it would be superfluous and useless to compute how surprising something was exactly. All the organism needs to know to derive respective action implications is the simple information "Surprise! Hold on!" or "No surprise, all fine, continue!", with more detailed analyses only requiring additional resources and potentially hindering effective reactions.

Yet, although I did not find significant differences between a medium and high degree of deviance in single comparisons, the descriptive patterns *do* largely point towards increasing response intensities with increasing deviance. Hence, the implication of dichotomous responses to surprise remains restricted to the current conservative testing within a very cognitive, highly controlled experimental paradigm – which might, however, be too coarse to capture the effects of subtler variations.

One of the most straightforward solutions to increase the probability of finding a true effect would be the usage of a within-subjects design. Already on a conceptual dimension, the implementation of different degrees of deviance implies a certain relativity (the perception of "medium deviance" may, for instance, be most accurate in the knowledge of what constitutes a high degree of deviance), which optimally operationally translates into all participants passing all conditions, ideally even with repeated measures for each cell. This would contribute to increasing the construct validity of graded deviance, allow for economically more efficient designs due to lower required sample sizes, and simultaneously decrease statistical error variance (e.g., Maxwell & Delaney, 2004). However, this option is undermined by the very

fundamental characteristics of surprise itself. By definition, at least perceptual surprise works only once as repeated encounters with perceptually unexpected deviations become increasingly expected, thereby successively reducing response strength (see Reisenzein et al., 2006, for empirical evidence). On these backgrounds, the present experiments employed a between-subjects design. In principle, though, a more fine-grained resolution of response strength differences could be achieved in two ways: On the one hand by triggering stronger responses, and on the other hand by increasing the measurement sensitivity. The stimuli employed in the present experiments were either uncontrolled, hence impeding any valid conclusions right from the outset, or highly impoverished, containing neither meaning nor provoking arousal nor holding significant motivational relevance. Boosting these factors could intensify the overall strength of the surprise responses and thereby foster a finer scaling of potential differences between varying degrees of deviance. Such a "boost" could be implemented by enhancing the goal relevance of the stimuli employed (for example by implementing a choice reaction task, e.g., Meyer et al., 1991; Niepel, Rudolph, Schützwohl, & Meyer, 1994), or by changing the set of stimuli in a way that increases involvement and emotional arousal (for example by adding meaning or social context). An alternative approach would be to induce a different type of surprise: Whereas the present research evoked *perceptual* surprise, surprise can also emerge on a *semantic* dimensions by disconfirming pre-existing knowledge structures – for instance, you would probably be surprised to learn that Vitamin C does not prevent a cold as you have taken this myth for granted until now (see also Reisenzein, 2000a; Topolinski & Strack, 2015). Since these pre-existing knowledge structures are more established and familiar than artificially prescribed expectations for a given experimental task, respective violations might have more profound effects and trigger stronger responses. Moreover, the induction of semantic surprise would also allow the implementation of an experimental within-subjects design, thereby presenting a viable alternative to enhance statistical power.

The second option to achieve more a more fine-grained resolution of the actual structure of surprise responses relates to increasing the sensitivity of the measurement method itself. This comprises amongst others the assessment of behavioral interruption via direct response time tasks (see the above-mentioned choice reaction tasks), or employing facial electromyography to measure subtle movements of affect-associated muscles, such as the *Musculus zygomaticus major* or the *Musculus corrugator supercilii* (e.g., Cacioppo, Petty, Losch & Kim, 1986; Topolinski, Likowski, Weyers & Strack, 2009; Scherer & Ellgring, 2007; see already Topolinski & Strack, 2015, for first promising investigations). Given that neurophysiological evidence on the dopaminergic coding of prediction errors implies that dopamine release or depression depend on the magnitude of the prediction error (e.g., Schultz, 1998), assessing the activity of dopamine neurons would likewise contribute to increasing our understanding of the structure of surprise responses. In that vein, the neurotransmitter serotonin proves equally interesting, as it seems to selectively carry information on the magnitude of a prediction error without simultaneously coding its valence (Matias et al., 2017).

Hence, the question on the graded-ness of the surprise responses cannot be conclusively answered yet, and the hunt for clarification continues.

4.1.3 On the Valence of Unpredictability – A Contextualization

Current theories on the valence of the unpredicted and the unpredictable cover almost the whole spectrum of perspectives, ranging from positivity over neutrality to negativity. Giving decisive input into these ongoing debates, the evidence reported in the current dissertation suggests a valence-lessness of the unpredicted and an averseness of the unpredictable.

That facing the unpredicted is essentially affect-free is in line with recent theories that conceptualize surprise as an innately neutral (pre-)cognitive state whose affective tone only results from the influence of contextual factors – this is, from analyzing whether the very event itself entails positive or negative consequences (Lazarus, 1991; Ludden et al., 2012; Mellers et

al., 2013; Ortony et al., 1988). Yet, the question that remains is: Was there indeed no affective response to surprise in the experiments reported in Chapter 2, or was this surprise-affect overwritten by the impact of contextual valence? If you see a green letter after a sequence of blue numbers, you will probably not care at all if I asked you about your preferences because those stimuli do not matter to you. Still, you can only verbalize what you are consciously aware of, and your immediate response to surprise "while it happens" might have been too short-lived to ever reach consciousness. On these backgrounds, the measurement of affective responses via explicit liking ratings presents one of the major flaws of the current research as it does not allow the undistorted assessment of on-line affective responses but is confounded with conscious evaluations. Future experiments may thus aim to unveil the "bare" valence of the unpredicted by achieving a more precise temporal resolution of the affective response dynamics, covering the entire processing range from stimulus onset to cognitive mastering.

That for the past – now let's turn to the future. In Chapter 3, I demonstrated a fundamental preference for the predictable in social interactions, with people being willing to forgo part of their potential returns to avoid interacting with an unpredictable partner. I quantified this psychological value of predictability with 31%: unless the unpredictable interaction option exceeds the predictable one by approximately this amount, people prefer choosing predictable interaction partners. This implies a negative valence of the unpredictable.

However, this is only the first-glance implication, and a thorough reply to the question on the valence of the unpredictable needs to go beyond the current paradigm and experiments. Firstly, the present estimate is restricted to social interactions, and we do not know whether and in how far non-social interactions (e.g., with bots) would trigger different patterns – a fruitful avenue for further research. Secondly, the experiments reported in Chapter 3 focused on assessing *relative* preferences, with participants having the choice between two options. Though, in our everyday lives, we cannot always choose between a predictable and an

unpredictable alternative: Receiving the job application of a candidate whose reference letters strongly vary in evaluations is, for example, not automatically escorted by the free addition of a consistently performing candidate's application but may stand on its own. The above-mentioned preliminary conclusion of a general preference for the predictable and an averseness of the unpredictable thus derives from calculating the *difference* in the valences of interacting with an unpredictable and a predictable partner – which, strictly speaking, tells us nothing about the absolute valence of the unpredictable per se. To be more precise, an observable preference for the predictable could in principle result from three possible scenarios: (1) Both the predictable and the unpredictable option are positively valenced, but the valence of the predictable option is somewhat more positive, (2) both the predictable and the unpredictable option are negatively valenced, but the valence of the predictable option is somewhat more positive, and (3) the predictable option is positively valenced and the unpredictable option is negatively valenced. Although the interpretation offered in Chapter 3 intuitively suggests the third scenario, a further unraveled look into the data in fact supports a fourth scenario, with preference ratings for the unpredictable interaction partner hardly deviating from the scale average of 5, implying neither positivity nor negativity, but neutrality. Since the present experiments were primarily designed to estimate the psychological *value* of predictability as an inherently relative construct (e.g., Kahneman & Tversky, 1979; von Neumann & Morgenstern, 1944; see also Ungemach, Stewart, & Reimers, 2011), an assessment of absolute preferences would have only decreased validity. Yet, future experiments that aim at unveiling the decontextualized valence of the unpredictable may do so by, for instance, implementing a between-subjects design with only one interaction partner per condition.

Thirdly, in a related vein, the valence of the unpredictable was currently derived from the psychological equivalent of the risk-return tradeoff. This implies already on a terminological dimension that I did not only consider the influence of risk (this is: the unpredictable), but likewise the influence of the expected return of an interaction. So, what

remains of the effect of (un)predictability when subtracting the impact of expected return? Eliminating the prospect of profit by asking participants to evaluate the mere likeability of their interaction partner (Chapter 3, Experiment 3) tentatively points into the direction of a certain attractiveness of the unpredictable. This suggests that situational goal orientation ("Are there potential gains for me?") can impact the hedonic qualities of the unpredictable.

Bundling the above-mentioned considerations thus implies that the valence of the unpredictable is largely shaped by the impact of context (see already Weber, Blais, & Betz, 2002, for evidence on the domain specificity of risk attitudes). Eliminating this context might, as provisional results connote, reveal an inherent valence-lessness. On the other hand, exploring this context opens a promising corridor for future research. While the presently reported experiments examine the hedonic properties of the (un)predictable when it comes to gaining potential work-related or financial payoff – and is hence located in a domain of risk aversion (see the *Prospect Theory*, Kahneman & Tversky, 1979) –, it would be equally interesting to expand these investigations to a loss domain in which people act more risk-seekingly. According to theory, the shift of preferences should occur earlier, and preferences might be even reversed; yet, to the best of my knowledge, no one ever put this case to test.

Further scientific endeavors may also address the potentially assimilative versus contrastive effects of context: Would our (un)predictability preferences adjust to the "defaults" of our surrounding environments and would we seek unpredictability if everything else is equally unpredictable vice versa (see also Maddi, 1968)? Or would we increasingly strive for predictability in an unpredictable world, and, in turn, favor unpredictability in an otherwise completely predictable world (see also McClelland et al., 1953)? In this light, Phil Connors, the already introduced time loop-caught protagonist of the movie "Groundhog Day", might also rather appreciate predictability if he was entrapped in an entirely unpredictable environment.

But this is how Phil might react, and we are not necessarily all like Phil. Perhaps it is only him who has the need for variation and escape from the time loop routine, and someone else would have savored a state of full predictability? Following this train of thought, context may, in the broadest sense, also refer to the "psychological milieu" in which the unpredicted or the unpredictable takes place – this is: to the characteristics of the individuals encountering unpredictability themselves. The research reported in Chapters 2 and 3 focused on the general, fixed effects of (un)predictability on affective responses and was not primarily interested in further interindividual differences. However, personality may shape the direction and intensity of these responses, and a sole consideration of event-related variables might hence not suffice to gain an in-depth understanding of the valence of unpredictability (see also Figner & Weber, 2015). As an exhaustive elaboration on the entirety of potentially involved personality variables would by far exceed the framework of this dissertation, I will at this point selectively focus on two variables that are most prominently discussed in the current unpredictability literature: *sensation seeking* and the *need for closure*.

Sensation seeking is defined as "the seeking of varied, novel, complex, and intense sensations and experiences, and the willingness to take physical, social, legal, and financial risks for the sake of such experience" (Zuckerman, 1994, p. 27). High sensation seekers thus strive for more thrill, adventure, and experience, are more inhibited and more susceptible to boredom than low sensation seekers (Zuckerman, 1971). Although the biological underpinnings of this construct have not been finally clarified yet, a growing number of findings points towards a crucial role of dopamine, with higher phasic and tonic dopaminergic activity in high compared to low sensation seekers, which results in different approach or avoidance tendencies towards high-stimulation stimuli (for a recent review, see Norbury & Husain, 2015; see also Zuckerman, 1990). This implies that attraction to and search for the unpredictable increase with increasing levels of sensation seeking (see also Horvath & Zuckerman, 1993).

The concept of need for closure denotes the desire for "an answer on a given topic, any answer, […] compared to confusion and ambiguity" (Kruglanski, 1990, p. 337). Individuals with a high need for closure experience a strong desire for predictability, a preference for order and structure, feel discomfort with ambiguity and have high levels of decisiveness and close-mindedness (Webster & Kruglanski, 1994). This need for closure does not necessarily comprise the strive for epistemic certainty in any domain, regardless of content (*nonspecific closure*), but may also refer to only very particular situations and questions (*specific closure*). Especially the associated desire for predictability points towards a role of this phenomenon in affective responses to unpredictability, suggesting that the unpredicted and the unpredictable become increasingly aversive with an increasing need for closure. In a related vein, Kruglanski et al. (2018) outline a theoretical perspective on affective responses to (in)consistency that builds on the assumption of an interplay between epistemic variables (i.e., updated expectancies about an outcome) and motivational variables (i.e., the subjective value attached to certainty on this outcome). With the latter being strongly impacted by the need for (specific and non-specific) closure, this perspective hence strongly implies that considering solely event-related variables, such as deviance or expectability, may cover only one side of the coin. Perhaps, there is in fact no such thing as *the* valence of unpredictability but a wealth of individual affective response patterns that simply average out when being aggregated. The decryption of this complex interplay between situational and personal variables, however, remains up to future research that empirically takes account of the very individual manifestation of these traits.

Speaking of context, I finally deem it necessary to point towards the notorious problem of controlled empirical research: ecological validity. The unpredicted and the unpredictable constitute phenomena which are solidly anchored in everyday lives, thereby actually calling for field experiments that go beyond computer screens, response buttons, and isolated cognitive paradigm bubbles. Such real-life investigations, however, conflict with the imperative of high experimental control. Yet, even though the scope for advancement remains broad and the top

of the range has been certainly not reached, the research presented in the current dissertation valuably contributes to successively increasing our understanding of unpredictability.

4.2 Conclusion

Addressing unsettled conflicts and cobbling hitherto loose threads together, the experiments presented in this dissertation enhanced our psychological understanding of the responses to and valence of the unpredicted and the unpredictable. To concludingly point out the key learnings of the current research: I demonstrated that responses to the unpredicted are driven by unexpectedness, with "expectations" referring to dynamically fine-tuning and recalibrating mental models. These responses seem insensitive to different gradations of surprising-ness and manifest in a dichotomous all-or-nothing manner. Beyond these processual insights, I highlighted that both the unpredicted and the unpredictable might be inherently valence-free and any affectively toned experience rather results from the hedonic impact of context. Despite these contributions, there still remains a vast field of implications to explore, and our current understanding may be far from exhaustive. Hence, the search for epistemic revelations shall go on.

5. References

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *The Journal of Neuroscience*, *30*(8), 2960–2966. https://doi.org/10.1523/JNEUROSCI.3730-10.2010

Arias-Carrión, Ó., & Pöppel, E. (2007). Dopamine, learning, and reward-seeking behavior. *Acta Neurobiologiae Experimentalis*, *67*(4), 481–488.

Arrow, K. J. (1965). *Aspects of the theory of risk bearing*. Helsinki: Yrjö Jahnsson Lectures.

Ashby, F. G., Isen, A. M., & Turken, A. U. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*, *106*(3), 529–550. https://doi.org/10.1037/0033-295x.106.3.529

Barnes, C. M., & Morgeson, F. P. (2007). Typical performance, maximal performance, and performance variability: Expanding our understanding of how organizations value performance. *Human Performance*, *20*(3), 259–274. https://doi.org/10.1080/08959280701333289

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. New York: McGraw-Hill Book Company. https://doi.org/10.1037/11164-000

Bernoulli, D. (1738/1954). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*(1), 23–36. https://doi.org/10.2307/1909829

Bodvarsson, Ö. B., & Brastow, R. T. (1998). Do employees pay for consistent performance?: Evidence from the NBA. *Economic Inquiry*, *36*(1), 145–160. https://doi.org/10.1111/j.1465-7295.1998.tb01702.x

Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review*, *98*(1), 294–310. https://doi.org/10.1257/aer.98.1.294

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, *55*(4), 467–484. https://doi.org/10.1016/j.jebo.2003.11.004

Braem, S., & Trapp, S. (2019). Humans show a higher preference for stimuli that are predictive relative to those that are predictable. *Psychological Research*, *83*(3), 567–573. https://doi.org/10.1007/s00426-017-0935-x

Brandenburger, A. M. (1996). Strategic and structural uncertainty in games. In R. J. Zeckhauser, R. L. Keeney, & J. K. Sebenius (Eds.), *Wise choices: Decisions, games, and negotiations* (pp. 221–232). Harvard: Business School Press.

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: Rewarding, aversive, and alerting. *Neuron*, *68*(5), 815–834. https://doi.org/10.1016/j.neuron.2010.11.022

Cacioppo, J. T., Petty, R. E., Losch, M. E., & Kim, H. S. (1986). Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology, 50*(2), 260–268. https://doi.org/10.1037//0022-3514.50.2.260

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. New York: Russell Sage Foundation.

Carlsmith, J. M., & Aronson, E. (1963). Some hedonic consequences of the confirmation and disconfirmation of expectancies. *Journal of Abnormal and Social Psychology, 66,* 151–156. https://doi.org/10.1037/h0042692

Carnaghi, A., Maass, A., Gresta, S., Bianchi, M., Cadinu, M., & Arcuri, L. (2008). Nomina sunt omina: On the inductive potential of nouns and adjectives in person perception. *Journal of Personality and Social Psychology*, *94*(5), 839–859. https://doi.org/10.1037/0022-3514.94.5.839

Chang, Y.-H., Levinboim, T., & Maheswaran, R. (2012). The Social Ultimatum Game. In T. V. Guy (Ed.), *Intelligent Systems Reference Library: Vol. 28. Decision making with imperfect decision-makers* (pp. 135–158). Berlin: Springer. https://doi.org/10.1007/978-3-642-24647-0_6

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, *17*(3), 521–534. https://doi.org/10.1007/s11097-017-9525-z

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. https://doi.org/10.1007/BF01320076

Coombs, C. H. (1975). Portfolio theory and the measurement of risk. In M. F. Kaplan, & S. Schwartz (Eds.), *Human judgment and decision* (pp. 63–85). New York: Academic Press.

Coull, J. T., & Nobre, A. C. (1998). Where and when to pay attention: The neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI. *The Journal of Neuroscience*, *18*(18), 7426–7435. https://doi.org/10.1523/JNEUROSCI.18-18-07426.1998

Cummings, R. G., Harrison, G. W., & Rutström, E. E. (1995). Homegrown values and hypothetical surveys: Is the dichotomous choice approach incentive-compatible? *American Economic Review*, *85*, 260–66.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London: John Murray.

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network*, *17*(4), 335–350. https://doi.org/10.1080/09548980601004024

de Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*, *7:10996*. https://doi.org/10.1038/ncomms10996

De Houwer, J., & Hermans, D. (1994). Differences in the affective processing of words and pictures. *Cognition & Emotion*, *8*(1), 1–20. https://doi.org/10.1080/02699939408408925

de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*, 1254–1258. https://doi.org/10.1126/science.1100735

DeNisi, A. S., & Stevens, G. E. (1981). Profiles of performance, performance evaluations, and personnel decisions. *Academy of Management Journal*, *24*(3), 592–602. https://doi.org/10.5465/255577

Deutscher, C., & Büschemann, A. (2016). Does performance consistency pay off financially for players? Evidence from the Bundesliga. *Journal of Sports Economics*, *17*(1), 27–43. https://doi.org/10.1177/1527002514521428

Deutscher, C., Gürtler, O., Prinz, J., & Weimar, D. (2017). The payoff to consistency in performance. *Economic Inquiry*, *55*(2), 1091–1103. https://doi.org/10.1111/ecin.12415

Diederen, K. M. J., & Fletcher, P. C. (2020). Dopamine, prediction error and beyond. *The Neuroscientist*, 1073858420907591. https://doi.org/10.1177/1073858420907591

Donchin, E. (1981). Surprise!... Surprise? *Psychophysiology*, *18*(5), 493–513. https://doi.org/10.1111/j.1469-8986.1981.tb01815.x

Ekman, P. (1972). Universals and cultural differences in facial expression of emotion. In. J. Cole (Ed.), *Nebraska symposium on motivation* (pp. 207–283). Lincoln: University of Nebraska Press.

Ekman, P. (1979). About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, & D. Ploog (Eds.), *Human ethology: Claims and limits of a new discipline* (pp. 169–202). Cambridge: University Press.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140. https://doi.org/10.1038/415137a

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00215

Ferrari, V., Bradley, M. M., Codispoti, M., & Lang, P. J. (2010). Detecting novelty and significance. *Journal of Cognitive Neuroscience, 22*(2), 404–411. https://doi.org/10.1162/jocn.2009.21244

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford: University Press.

Fetchenhauer, D., & Dunning, D. (2012). Betrayal aversion versus principled trustfulness – How to explain risk avoidance and risky choices in trust games. *Journal of Economic Behavior & Organization*, *81*(2), 534–541. https://doi.org/10.1016/j.jebo.2011.07.017

Figner, B., & Weber, E. U. (2015). Personality and risk taking. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 809–813). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.26047-9

Fiorillo, C. D. (2013). Two dimensions of value: Dopamine neurons represent reward but not aversiveness. *Science*, *341*(6145), 546–549. https://doi.org/10.1126/science.1238699

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–1902. https://doi.org/10.1126/science.1077349

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 552–564. https://doi.org/10.1037/0096-1523.3.4.552

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, *18*(12), 1050–1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x

Foster, M. I., & Keane, M. T. (2015). Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty. *Cognitive Psychology, 81*, 74–116. https://doi.org/10.1016/j.cogpsych.2015.08.004

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances, 3*(10), e1701381. https://doi.org/10.1126/sciadv.1701381

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364*(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology, 3*, 130. https://doi.org/10.3389/fpsyg.2012.00130

Furtner, M. R., Rauthmann, J. F., & Sachse, P. (2009). Nomen est omen: Investigating the dominance of nouns in word comprehension with eye movement analyses. *Advances in Cognitive Psychology, 5*, 91–104. https://doi.org/10.2478/v10053-008-0069-1

Gabay, A. S., Radua, J., Kempton, M. J., & Mehta, M. A. (2014). The Ultimatum Game and the brain: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews, 47*, 549–558. https://doi.org/10.1016/j.neubiorev.2014.10.014

Gawronski, B., & Brannon, S. M. (2019). What is cognitive consistency, and why does it matter? In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a pivotal theory in psychology* (pp. 91–116). Washington: American Psychological Association. https://doi.org/10.1037/0000135-005

Gawronski, B., & Strack, F. (2012). Cognitive consistency as a basic principle of social information processing. In B. Gawronski & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 1–16). New York: Guilford Press.

Ghysels, Santa-Clara, P., & Valkanov, R. (2005). There is a risk-return trade-off after all. *Journal of Financial Economics*, *76*(3), 509–548. https://doi.org/10.1016/j.jfineco.2004.03.008

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science*, *317*(5843), 1351–1354. https://doi.org/10.1126/science.1144161

Grace, A. A. (1991). Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: A hypothesis for the etiology of schizophrenia. *Neuroscience*, *41*(1), 1–24. https://doi.org/10.1016/0306-4522(91)90196-U

Grillon, C., Baas, J. P., Lissek, S., Smith, K., & Milstein, J. (2004). Anxious responses to predictable and unpredictable aversive events. *Behavioral Neuroscience*, *118*(5), 916–924. https://doi.org/10.1037/0735-7044.118.5.916

Güth, W., & Kocher, M. G. (2014). More than thirty years of Ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization*, *108*, 396–409. https://doi.org/10.1016/j.jebo.2014.06.006

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of Ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Hajcak, G., Holroyd, C. B., Moser, J. S., & Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology*, *42*(2), 161–170. https://doi.org/10.1111/j.1469-8986.2005.00278.x

Harmon-Jones, E., Amodio, D. M., & Harmon-Jones, C. (2009). Action-based model of dissonance: A review, integration, and expansion of conceptions of cognitive conflict. In M. P. Zanna (Ed.), *Advances in experimental social psychology, Vol. 41* (pp. 119–166). Burlington: Elsevier Academic Press.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: John Wiley and Sons.

Heine, S. J., Proulx, T., & Vohs, K. D. (2006). The meaning maintenance model: On the coherence of social motivations. *Personality and Social Psychology Review, 10*(2), 88–110. https://doi.org/10.1207/s15327957pspr1002_1

Heinemann, F., Nagel, R., & Ockenfels, P. (2009). Measuring strategic uncertainty in coordination games. *Review of Economic Studies, 76*(1), 181–221. https://doi.org/10.1111/j.1467-937X.2008.00512.x

Herry, C., Bach, D. R., Esposito, F., Di Salle, F., Perrig, W. J., Scheffler, K., Lüthi, A., & Seifritz, E. (2007). Processing of temporal unpredictability in human and animal amygdala. *The Journal of Neuroscience, 27*(22), 5958–5966. https://doi.org/10.1523/JNEUROSCI.5218-06.2007

Hinds, P. J., Carley, K.M., Krackhardt, D., & Wholey, D. (2000). Choosing work group members: Balancing similarity, competence, and familiarity. *Organizational Behavior and Human Decision Processes, 81*(2), 226–251. https://doi.org/10.1006/obhd.1999.2875

Hinojosa, J. A., Carretié, L., Valcárcel, M. A., Méndez-Bértolo, C., & Pozo, M. A. (2009). Electrophysiological differences in the processing of affective information in words and pictures. *Cognitive, Affective & Behavioral Neuroscience, 9*(2), 173–189. https://doi.org/10.3758/CABN.9.2.173

Hirsh, J. B., Mar, R. A., & Peterson, J. B. (2012). Psychological entropy: A framework for understanding uncertainty-related anxiety. *Psychological Review*, *119*(2), 304–320. https://doi.org/10.1037/a0026767

Hogg, M. A. (2000). Subjective uncertainty reduction through self-categorization: A motivational theory of social identity processes. *European Review of Social Psychology*, *11*(1), 223–255. https://doi.org/10.1080/14792772043000040

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3*, 96. https://doi.org/10.3389/fpsyg.2012.00096

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *Neuroreport*, *14*(18), 2481–2484. https://doi.org/10.1097/01.wnr.0000099601.41403.a5

Horstmann, G. (2002). Evidence for attentional capture by a surprising color singleton in visual search. *Psychological Science*, *13*(6), 499–505. https://doi.org/10.1111/1467-9280.00488

Horstmann, G. (2005). Attentional capture by an unannounced color singleton depends on expectation discrepancy. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(5), 1039–1060. https://doi.org/10.1037/0096-1523.31.5.1039

Horstmann, G. (2006). Latency and duration of the action interruption in surprise. *Cognition & Emotion*, *20*(2), 242–273. https://doi.org/10.1080/02699930500262878

Horstmann, G. (2015). The surprise-attention link: A review. *Annals of the New York Academy of Sciences*, *1339*(1), 106–115. https://doi.org/10.1111/nyas.12679

Horstmann, G., & Schützwohl, A. (1998). Zum Einfluss der Verknüpfungsstärke von Schemaelementen auf die Stärke der Überraschungsreaktion. *Zeitschrift für Experimentelle Psychologie*, *45*, 203–217.

Horvath, P., & Zuckerman, M. (1993). Sensation seeking, risk appraisal, and risky behavior. *Personality and Individual Differences*, *14*(1), 41–52. https://doi.org/10.1016/0191-8869(93)90173-Z

IBM Corp. (2017). *IBM SPSS Statistics for Windows* (Version 25.0) [Computer Software]. Armonk, NY: IBM Corp.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306. https://doi.org/10.1016/j.visres.2008.09.007

Izard, C. E. (1971). *The face of emotion*. East Norwalk: Appleton-Century-Crofts.

Jackson, F., Nelson, B. D., & Proudfit, G. H. (2015). In an uncertain world, errors are more aversive: Evidence from the error-related negativity. *Emotion*, *15*(1), 12–16. https://doi.org/10.1037/emo0000020

Jarvis, B. G. (2014). *DirectRT* (Version 2014.1.123) [Computer Software]. New York: Empirisoft Corporation.

Johannesson, M., Liljas, B., & Johansson, P.-O. (1998). An experimental comparison of dichotomous choice contingent valuation questions and real purchase decisions. *Applied Economics*, *30*(5), 643–647. https://doi.org/10.1080/000368498325633

Kagan, J. (1972). Motives and development. *Journal of Personality and Social Psychology*, *22*(1), 51–66. https://doi.org/10.1037/h0032356

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153. https://doi.org/10.1037/0033-295X.93.2.136

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263. https://doi.org/10.2307/1914185

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*(2), 143–157. https://doi.org/10.1016/0010-0277(82)90023-3

Kelly, E. L. (1955). Consistency of the adult personality. *American Psychologist, 10*(11), 659–681. https://doi.org/10.1037/h0040747

Kensinger, E.A., & Schacter, D.L. (2006). Processing emotional pictures and words: Effects of valence and arousal. *Cognitive, Affective & Behavioral Neuroscience, 6*(2), 110–126. https://doi.org/10.3758/CABN.6.2.110

Knight, F. H. (1921). *Risk, uncertainty, and profit*. Boston: Houghton Mifflin.

Kruglanski, A. W. (1990). Motivations for judging and knowing: Implications for causal attribution. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior, Vol. 2* (pp. 333–368). New York: The Guilford Press.

Kruglanski, A. W., Jasko, K., Milyavsky, M., Chernikova, M., Webber, D., Pierro, A., & Di Santo, D. (2018). Cognitive consistency theory in social psychology: A paradigm reconsidered. *Psychological Inquiry*, *29*(2), 45–59. https://doi.org/10.1080/1047840X.2018.1480619

Lambooij, M. S., Harmsen, I. A., Veldwijk, J., Melker, H. de, Mollema, L., van Weert, Y. W. M., & de Wit, G. A.. (2015). Consistency between stated and revealed preferences: A discrete choice experiment and a behavioural experiment on vaccination behaviour compared. *BMC Medical Research Methodology*, *15*(1). https://doi.org/10.1186/s12874-015-0010-5

Larsen, J. T., Norris, C. J., & Cacioppo, J. T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, *40*(5), 776–785. https://doi.org/10.1111/1469-8986.00078

Lauffs, M. M., Geoghan, S. A., Favrod, O., Herzog, M. H., & Preuschoff, K. (2020). Risk prediction error signaling: A two-component response? *NeuroImage*, *214*, 116766. https://doi.org/10.1016/j.neuroimage.2020.116766

Lazarus, R. S. (1991). *Emotion and adaptation.* Oxford: University Press.

Levy, N., Harmon-Jones, C., & Harmon-Jones, E. (2018). Dissonance and discomfort: Does a simple cognitive inconsistency evoke a negative affective state? *Motivation Science*, *4*(2), 95–108. https://dx.doi.org/10.1037/mot0000079

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, *67*(1), 145–163.

Loewenstein, J. (2019). Surprise, recipes for surprise, and social influence. *Topics in Cognitive Science*, *11*(1), 178–193. https://doi.org/10.1111/tops.12312

Lorini, E., & Castelfranchi, C. (2007). The cognitive structure of surprise: Looking for basic principles. *Topoi*, *26*(1), 133–149. https://doi.org/10.1007/s11245-006-9000-x

Ludden, G., Schifferstein, R., & Hekkert, P. (2012). Beyond Surprise: A longitudinal study of responses to visual-tactual incongruities in products. *International Journal of Design*, *6*, 1–10. Macedo, L., & Cardoso, A. (2017).

Macedo, L., & Cardoso, A. (2019). A contrast-based computational model of surprise and its applications. *Topics in Cognitive Science*, *11*(1), 88–102. https://doi.org/10.1111/tops.12310

Macedo, L., Cardoso, A., Reisenzein, R**.**, Lorini, L., & Castelfranchi, C. (2009). Artificial surprise. In J. Vallverdú, & D. Casacuberta (Eds.) *Handbook of research on synthetic emotions and sociable robotics: New applications in affective computing and artificial intelligence* (pp. 267-291*).* Hershey: IGI Global.

Maddi, S. A. (1968). The pursuit of consistency and variety. In R.P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J.Rosenberg, & P. H. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook* (pp. 267–274). Chicago: RandMcNally.

Maguire, R., & Keane, M. T. (2006). Surprise: Disconfirmed expectations or representation-fit? In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1765–1770). Hillsdale: Erlbaum.

Maguire, R., Maguire, P., & Keane, M. T. (2011). Making sense of surprise: An investigation of the factors influencing surprise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 176–186. https://doi.org/10.1037/a0021609

Mandler, G. (1975). *Mind and emotion*. New York: Wiley.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

Mata, R., Frey, R., Richter, D., Schupp, J., & Hertwig, R. (2018). Risk preference: A view from psychology. *Journal of Economic Perspectives*, *32*(2), 155–172. https://doi.org/10.1257/jep.32.2.155

Matias, S., Lottem, E., Dugué, G. P., & Mainen, Z. F. (2017). Activity patterns of serotonin neurons underlying cognitive flexibility. *ELife*, *6*. https://doi.org/10.7554/eLife.20552

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective* (2nd ed). Mahwah: Lawrence Erlbaum Associates.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive*. East Norwalk: Appleton-Century-Crofts. https://doi.org/10.1037/11144-000

Mellers, B., Fincher, K., Drummond, C., & Bigony, M. (2013). Surprise: A belief or an emotion? *Progress in Brain Research*, *202*, 3–19. https://doi.org/10.1016/B978-0-444-62604-2.00001-0

Mellers, B., Schwartz, A., & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General*, *128*(3), 332–345. https://doi.org/10.1037/0096-3445.128.3.332

Mendes, W. B., Blascovich, J., Hunter, S. B., Lickel, B., & Jost, J. T. (2007). Threatened by the unexpected: Physiological responses during social interactions with expectancy-violating partners. *Journal of Personality and Social Psychology*, *92*(4), 698–716. https://doi.org/10.1037/0022-3514.92.4.698

Merton, R. C. (1980). On estimating the expected return on the market. *Journal of Financial Economics*, *8*(4), 323–361. https://doi.org/10.1016/0304-405X(80)90007-0

Meyer, W.-U., Niepel, M., Rudolph, U., & Schützwohl, A. (1991). An experimental analysis of surprise. *Cognition & Emotion*, *5*(4), 295–311. https://doi.org/10.1080/02699939108411042

Meyer, W.-U., Reisenzein, R. & Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. *Motivation & Emotion*, 21(3), 251–274. https://doi.org/10.1023/A:1024422330338

Miceli, M., & Castelfranchi, C. (2002). The mind and the future. *Theory & Psychology*, *12*(3), 335–366. https://doi.org/10.1177/0959354302012003015

Montgomery, H., Selart, M., Gärling, T., & Lindberg, E. (1994). The judgment-choice discrepancy: Noncompatibility or restructuring? *Journal of Behavioral Decision Making*, *7*(2), 145–155. https://doi.org/10.1002/bdm.3960070207

Munnich, E. L., Foster, M. I., & Keane, M. T. (2019). Editors' introduction and review: An appraisal of surprise: Tracing the threads that stitch it together. *Topics in Cognitive Science*, *11*(1), 37–49. https://doi.org/10.1111/tops.12402

Neta, M., Norris, C. J., & Whalen, P. J. (2009). Corrugator muscle responses are associated with individual differences in positivity-negativity bias. *Emotion*, *9*(5), 640–648. https://doi.org/10.1037/a0016819

Newman, J., Krzystofiak, F., & Cardy, R. (1986). Role of behavior level, behavioral variability, and ratee order in the formation of appraisal ratings. *Basic and Applied Social Psychology*, *7*(4), 277–293. https://doi.org/10.1207/s15324834basp0704_3

Niepel, M. (2001). Independent manipulation of stimulus change and unexpectedness dissociates indices of the orienting response. *Psychophysiology*, *38*(1), 84–91. https://doi.org/10.1111/1469-8986.3810084

Niepel, M., Rudolph, U., Schützwohl, A., & Meyer, W.-U. (1994). Temporal characteristics of the surprise reaction induced by schema-discrepant visual and auditory events. *Cognition & Emotion*, *8*(5), 433–452. https://doi.org/10.1080/02699939408408951

Noordewier, M. K., & Breugelmans, S. M. (2013). On the valence of surprise. *Cognition & Emotion*, *27*(7), 1326–1334. https://doi.org/10.1080/02699931.2013.777660

Noordewier, M. K., Topolinski, S., & Van Dijk, E. (2016). The temporal dynamics of surprise. *Social and Personality Psychology Compass*, *10*(3), 136–149. https://doi.org/10.1111/spc3.12242

Noordewier, M. K., & Van Dijk, E. (2018). Surprise: Unfolding of facial expressions. *Cognition & Emotion*, 1–16. https://doi.org/10.1080/02699931.2018.1517730

Norbury, A., & Husain, M. (2015). Sensation-seeking: Dopaminergic modulation and risk for psychopathology. *Behavioural Brain Research*, *288*, 79–93. https://doi.org/10.1016/j.bbr.2015.04.015

Ogawa, H., & Watanabe, K. (2011). Implicit learning increases preference for predictive visual display. *Attention, Perception & Psychophysics*, *73*(6), 1815–1822. https://doi.org/10.3758/s13414-010-0041-2

Oosterbeek, H., Sloof, R., & van de Kuilen, G. (2004). Cultural differences in Ultimatum Game experiments: Evidence from a meta-analysis. *Experimental Economics*, *7*, 171–188. https://doi.org/10.1023/B:EXEC.0000026978.14316.74

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions.* Cambridge: University Press. https://doi.org/10.1017/CBO9780511571299

Ortony, A., & Partridge, D. (1987). Surprisingness and expectation failure: What's the difference? *Proceedings of the 10th International Joint Conference on Artificial Intelligence* (pp. 106–108). Los Altos: Morgan Kaufmann.

Peters, A., McEwen, B. S., & Friston, K. (2017). Uncertainty and stress: Why it causes diseases and how it is mastered by the brain. *Progress in Neurobiology*, *156*, 164–188. https://doi.org/10.1016/j.pneurobio.2017.05.004

Pezzo, M. V. (2003). Surprise, defence, or making sense: What removes hindsight bias? *Memory*, *11*(4-5), 421–441. https://doi.org/10.1080/09658210244000603

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. New York: Harper & Row.

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*(2), 160–174. https://doi.org/10.1037/0096-3445.109.2.160

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, *31*(1–2), 122–136.

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, *51*(3), 381–390. https://doi.org/10.1016/j.neuron.2006.06.024

Proulx, T., & Heine, S. J. (2008). The case of the transmogrifying experimenter: Affirmation of a moral schema following implicit change detection. *Psychological Science*, *19*(12), 1294–1300. https://doi.org/10.1111/j.1467-9280.2008.02238.x

Proulx, T., Heine, S. J., & Vohs, K. D. (2010). When is the unfamiliar the uncanny? Meaning affirmation after exposure to absurdist literature, humor, and art. *Personality & Social Psychology Bulletin*, *36*(6), 817–829. https://doi.org/10.1177/0146167210369896

Proulx, T., & Inzlicht, M. (2012). The five "A"s of meaning maintenance: Finding meaning in the theories of sense-making. *Psychological Inquiry*, *23*(4), 317–335. https://doi.org/10.1080/1047840X.2012.702372

Proulx, T., Inzlicht, M., & Harmon-Jones, E. (2012). Understanding all inconsistency compensation as a palliative response to violated expectations. *Trends in Cognitive Sciences*, *16*(5), 285–291. https://doi.org/10.1016/j.tics.2012.04.002

Proulx, T., Sleegers, W., & Tritt, S. M. (2017). The expectancy bias: Expectancy-violating faces evoke earlier pupillary dilation than neutral or negative faces. *Journal of Experimental Social Psychology*, *70*, 69–79. https://doi.org/10.1016/j.jesp.2016.12.003

Przysinda, E., Zeng, T., Maves, K., Arkin, C., & Loui, P. (2017). Jazz musicians reveal role of expectancy in human creativity. *Brain and Cognition*, *119*, 45–53. https://doi.org/10.1016/j.bandc.2017.09.008

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Reisenzein, R. (2000a). Exploring the strength of association between the components of emotion syndromes: The case of surprise. *Cognition & Emotion*, *14*(1), 1–38. https://doi.org/10.1080/026999300378978

Reisenzein, R. (2000b). The subjective experience of surprise. In H. Bless & J. P. Forgas (Eds.), *The message within: The role of subjective experience in social cognition and behavior* (pp. 262–279). New York: Psychology Press.

Reisenzein, R., Bördgen, S., Holtbernd, T., & Matz, D. (2006). Evidence for strong dissociation between emotion and facial displays: The case of surprise. *Journal of Personality and Social Psychology*, *91*(2), 295–315. https://doi.org/10.1037/0022-3514.91.2.295

Reisenzein, R., Horstmann, G., & Schützwohl, A. (2019). The cognitive-evolutionary model of surprise: A review of the evidence. *Topics in Cognitive Science*, *11*(1), 50–74. https://doi.org/10.1111/tops.12292

Reisenzein, R., Meyer, W.-U., & Niepel, M. (2012). Surprise. In V. S. Ramachandran (Ed.), *Encyclopedia of Human Behavior, 2nd ed*. (pp. 564–570). London: Elsevier Academic Press.

Reisenzein, R., & Studtmann, M. (2007). On the expression and experience of surprise: No evidence for facial feedback, but evidence for a reverse self-inference effect. *Emotion*, *7*(3), 612–627. https://doi.org/10.1037/1528-3542.7.3.612

Reisenzein, R., Studtmann, M., & Horstmann, G. (2013). Coherence between emotion and facial expression: Evidence from laboratory experiments. *Emotion Review, 5*(1), 16–23. https://doi.org/10.1177/1754073912457228

Retell, J. D., Becker, S. I., & Remington, R. W. (2016). Previously seen and expected stimuli elicit surprise in the context of visual search. *Attention, Perception & Psychophysics, 78*(3), 774–788. https://doi.org/10.3758/s13414-015-1052-9

Rieger, M. O., Wang, M., & Hens, T. (2014). Risk preferences around the world. *Management Science, 61*(3), 637-648.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Ruch, W. (1995). Will the real relationship between facial expression and affective experience please stand up: The case of exhilaration. *Cognition & Emotion, 9*(1), 33–58. https://doi.org/10.1080/02699939508408964

Rumelhart, D. E., & Norman, D. A. (1978). Accretion, tuning, and restructuring: Three modes of learning. In J.W. Cotton & R. Klatzky (Eds.), *Semantic factors in cognition* (pp. 37–53). Hillsdale: Erlbaum.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–135). Hillsdale: Erlbaum.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*, 1755–1758. https://doi.org/10.1126/science.1082976

Sarinopoulos, I., Grupe, D. W., Mackiewicz, K. L., Herrington, J. D., Lor, M., Steege, E. E., & Nitschke, J. B. (2010). Uncertainty during anticipation modulates neural responses to aversion in human insula and amygdala. *Cerebral Cortex*, *20*(4), 929–940. https://doi.org/10.1093/cercor/bhp155

Scherer, K. R., & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, *7*(1), 113–130. https://doi.org/10.1037/1528-3542.7.1.113

Scherer, K. R., Zentner, M. R., & Stern, D. (2004). Beyond surprise: The puzzle of infants' expressive reactions to expectancy violation. *Emotion*, *4*(4), 389–402. https://doi.org/10.1037/1528-3542.4.4.389

Schiffer, A.-M., Ahlheim, C., Wurm, M. F., & Schubotz, R. I. (2012). Surprised at all the entropy: Hippocampal, caudate and midbrain contributions to learning from prediction errors. *PloS One*, *7*(5), e36445. https://doi.org/10.1371/journal.pone.0036445

Schmidt, D., Shupp, R., Walker, J. M., & Ostrom, E. (2003). Playing safe in coordination games. *Games and Economic Behavior*, *42*(2), 281–299. https://doi.org/10.1016/S0899-8256(02)00552-3

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schubert, L., Körner, A., Lindau, B., Strack, F., & Topolinski, S. (2017). Open-minded midwifes, literate butchers, and greedy hooligans – The independent contributions of stereotype valence and consistency on evaluative judgments. *Frontiers in Psychology, 8:1723*. https://doi.org/10.3389/fpsyg.2017.01723

Schultz, W. (1998). Predictive reward signal of dopamine neurons. Journal of *Neurophysiology*, *80*(1), 1–27. https://doi.org/10.1152/jn.1998.80.1.1

Schultz, W. (2016). Dopamine reward prediction-error signalling: A two-component response. *Nature Reviews Neuroscience*, *17*(3), 183–195. https://doi.org/10.1038/nrn.2015.26

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience, 23*, 473–500. https://doi.org/10.1146/annurev.neuro.23.1.473

Schützwohl, A. (1998). Surprise and schema strength. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(5), 1182–1199. https://doi.org/10.1037/0278-7393.24.5.1182

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425–442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

Simandan, D. (2018). Being surprised and surprising ourselves. *Progress in Human Geography*, *64*(3), 030913251881043. https://doi.org/10.1177/0309132518810431

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99. https://doi.org/10.2307/1884852

Sinaceur, M., Adam, H., van Kleef, G. A., & Galinsky, A. D. (2013). The advantages of being unpredictable: How emotional inconsistency extracts concessions in negotiation. *Journal of Experimental Social Psychology*, *49*(3), 498–508. https://doi.org/10.1016/j.jesp.2013.01.007

Smedslund, J. (1990). Psychology and psychologic: Characterization of the difference. In K. J. Gergen & G. R. Semin (Eds.), *Everyday understanding: Social and scientific implications* (Vol. 29, pp. 45–63). London: Sage.

Steinel, W., van Beest, I., & van Dijk, E. (2014). Too good to be true: Suspicion-based rejections of high offers. *Group Processes & Intergroup Relations*, *17*(5), 682–698. https://doi.org/10.1177/1368430213507323

Stiensmeier-Pelster, J., Martini, A., & Reisenzein, R. (1995). The role of surprise in the attribution process. *Cognition & Emotion*, *9*(1), 5–31. https://doi.org/10.1080/02699939508408963

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., & Friston, K. J. (2005). Information theory, novelty and hippocampal responses: Unpredicted or unpredictable? *Neural Networks*, *18*(3), 225–230. https://doi.org/10.1016/j.neunet.2004.12.004

Suls, J. M. (1972). A two-stage model for the appreciation of jokes and cartoons. In J. H. Goldstein and P. E. McGhee (Eds.), *The psychology of humor: Theoretical perspectives and empirical issues* (pp. 81–100). New York: Academic Press.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D.,. . . Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*(2), 247–250. https://doi.org/10.1016/j.jml.2004.03.002

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*, *4*, 250. https://doi.org/10.3389/fpsyg.2013.00250

Teigen, K. H., & Keren, G. (2003). Surprises: Low probabilities or high contrasts? *Cognition*, *87*(2), 55–71. https://doi.org/10.1016/s0010-0277(02)00201-9

Thomaschke, R., & Dreisbach, G. (2013). Temporal predictability facilitates action, not perception. *Psychological Science*, *24*(7), 1335–1340. https://doi.org/10.1177/0956797612469411

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, *307*(5715), 1642–1645. https://doi.org/10.1126/science.1105370

Tobler, P. N., O'doherty, J. P., Dolan, R. J., & Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology*, *95*(1), 301–310. https://doi.org/10.1152/jn.00762.2005

Topolinski, S., Likowski, K. U., Weyers, P., & Strack, F. (2009). The face of fluency: Semantic coherence automatically elicits a specific pattern of facial muscle reactions. *Cognition & Emotion*, *23*(2), 260–271. https://doi.org/10.1080/02699930801994112

Topolinski, S., Maschmann, I. T., Pecher, D., & Winkielman, P. (2014). Oral approach–avoidance: Affective consequences of muscular articulation dynamics. *Journal of Personality and Social Psychology*, *106*(6), 885–896. https://doi.org/10.1037/a0036477

Topolinski, S., & Strack, F. (2009). Motormouth: Mere exposure depends on stimulus-specific motor simulations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 423–433. https://doi.org/10.1037/a0014504

Topolinski, S., & Strack, F. (2015). Corrugator activity confirms immediate negative affect in surprise. *Frontiers in Psychology, 6.* https://doi.org/10.3389/fpsyg.2015.00134

Trapp, S., Shenhav, A., Bitzer, S., & Bar, M. (2015). Human preferences are biased towards associative information. *Cognition & Emotion*, *29*(6), 1054–1068. https://doi.org/10.1080/02699931.2014.966064

Tversky, B. (1969). Pictorial and verbal encoding in a short-term memory task. *Perception & Psychophysics*, *6*(4), 225–233. https://doi.org/10.3758/BF03207022

Tversky, A., & Griffin, D. (1991). Endowment and contrast in judgments of well-being. In F. Strack, M. Argyle, & N. Schwarz (Eds.), *International series in experimental social psychology, Vol. 21. Subjective well-being: An interdisciplinary perspective* (pp. 101-118). Elmsford: Pergamon Press.

Underwood, B. J. (1952). An orientation for research on thinking. *Psychological Review*, *59*(3), 209–220. https://doi.org/10.1037/h0059415

Ungemach, C., Stewart, N., & Reimers, S. (2011). How incidental values from the environment affect decisions about money, risk, and delay. *Psychological Science*, *22*(2), 253–260. https://doi.org/10.1177/0956797610396225

Vanhamme, J. (2003). Surprise . . . surprise. An empirical investigation on how surprise is connected to consumer satisfaction. *ERIM Report Series Research in Management ERS-2003–005-MKT*. Retrieved from http://repub.eur.nl/res/pub/273/erimrs20030211172951.pdf

Veatch, T. C. (1998). A theory of humor. *Humor: International Journal of Humor Research, 11*(2), 161–215. https://doi.org/10.1515/humr.1998.11.2.161

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: University Press.

Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making, 15*(4), 263–290. https://doi.org/10.1002/bdm.414

Weber, E. U., & Hsee, C. (1998). Cross-cultural differences in risk perception, but cross-cultural similarities in attitudes towards perceived risk. *Management Science, 44*(9), 1205-1217.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67*(6), 1049–1062. https://doi.org/10.1037/0022-3514.67.6.1049

Wessel, J. R., Jenkinson, N., Brittain, J.-S., Voets, S. H. E. M., Aziz, T. Z., & Aron, A. R. (2016). Surprise disrupts cognition via a fronto-basal ganglia suppressive mechanism. *Nature Communications, 7*, 11195. https://doi.org/10.1038/ncomms11195

Wickelgren, I. (1997). Getting the brain's attention. *Science, 278*(5335), 35–37. https://doi.org/10.1126/science.278.5335.35

Wilson, T. D., Centerbar, D. B., Kermer, D. A., & Gilbert, D. T. (2005). The pleasures of uncertainty: Prolonging positive moods in ways people do not anticipate. *Journal of Personality and Social Psychology, 88*(1), 5–21. https://doi.org/10.1037/0022-3514.88.1.5

Wilson, T. D., & Gilbert, D. T. (2008). Explaining away: A model of affective adaptation. *Perspectives on Psychological Science*, *3*(5), 370–386. https://doi.org/10.1111/j.1745-6924.2008.00085.x

Wise, R. A. (1980). The dopamine synapse and the notion of 'pleasure centers' in the brain. *Trends in Neurosciences*, *3*(4), 91–95. https://doi.org/10.1016/0166-2236(80)90035-1

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, *5*(6), 483–494. https://doi.org/10.1038/nrn1406

Wise, R. A. (2008). Dopamine and reward: The anhedonia hypothesis 30 years on. *Neurotoxicity Research*, *14*(2-3), 169–183. https://doi.org/10.1007/BF03033808

Worthen, J. B., Coats, S., McGlynn, R. P., & Rossano, M. J. (2007). Cognitive factors in the prediction of liking of social groups: Prototypes, predictability and familiarity. In J. A. Zebowski (Ed.), *New research on social perception* (pp. 161–179). Hauppauge: Nova Science Publishers.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Zuckerman, M. (1971). Dimensions of sensation seeking. *Journal of Consulting and Clinical Psychology*, *36*(1), 45–52. https://doi.org/10.1037/h0030478

Zuckerman, M. (1990). The psychophysiology of sensation seeking. *Journal of Personality*, *58*(1), 313–345. https://doi.org/10.1111/j.1467-6494.1990.tb00918.x

Zuckerman, M. (1994). *Behavioral expressions and biosocial bases of sensation seeking*. Cambridge: University Press.