Multilevel Reward Prediction Errors, Not Expectations Or Outcomes, Drive Emotional Valence

Thalia H. Vrantsidis[1], Alan Voodla[2,3], Kimia Sabbagh[1]

1. Department of Psychology, Mississippi State University, Mississippi, United States

2. Institute of Psychology, University of Tartu, Tartu, Estonia

3. Brain and Cognition, KU Leuven, Leuven, Belgium

This is a pre-print, version 3. This paper has not been peer reviewed.

Author Note

The pre-registration, materials, data, and analysis scripts can be found here: https://osf.io/v5ubg/?view_only=bf3e05ef57f74e988a9941892caadf18

This research was supported by funding from the Estonian Ministry of Education and Research, which funded the Estonian Center of Excellence of Well-Being Sciences (grant number TK218). Correspondence concerning this article should be addressed to Thalia Vrantsidis, Department of Psychology, Rice Hall, Mississippi State University, Mississippi State, MS 39762. ORCID: 0000-0003-0766-9041. Email: tvrantsidis@psychology.msstate.edu

Author Contributions Statement

Thalia Vrantsidis: Conceptualization, Formal analysis, Methodology, Software, Supervision, Visualization, Writing—original draft, Writing—review and editing. Lead contributor in all listed roles. **Alan Voodla**: Conceptualization, Formal analysis, Methodology, Software, Writing—original draft, Writing—review and editing. Supporting contributor in all listed roles. **Kimia Sabbagh**: Methodology, Writing—original draft, Writing—review and editing. Supporting contributor in all listed roles.

Abstract

What role do expectations, outcomes, and their mismatches—i.e., reward prediction errors (RPEs)—play in generating emotions? Existing theories make competing predictions, with value-based perspectives suggesting that emotional valence should track outcomes and expectations, value-updating perspectives suggesting that it should instead track RPEs, and other views suggesting a role for all three factors. Yet empirical tests of these predictions have produced mixed results. Here, we demonstrate how these conflicting results could be reconciled within a value-updating perspective, by considering multiple forms of RPEs and their impacts over time. In particular, we build off recent work which suggested that affect in a perceptual decision-making task was influenced by both expectations and outcomes, but not by RPEs. The current study re-evaluates this conclusion, by modifying the original task and analyses to account for both trial-level and block-level RPEs and their temporal dynamics. With these modifications, results showed that both forms of RPEs drove affect in this task, with no clear effect of expectations or outcomes beyond this, contrary to previous conclusions. This work supports the central role of RPEs in driving emotional valence, in line with value-updating perspectives. Moreover, linking emotions to RPEs and value-updating has broad theoretical implications, such as highlighting how resolving prediction errors may be key to resolving emotions, and how difficulties in this process may underlie many emotion-related disorders.

Keywords: emotion, affect, prediction error, expectation, reinforcement learning

Introduction

What brings about positive or negative emotions? Things being good or bad? Expecting things to be good or bad? Things being better or worse than one's expectations? Or perhaps a combination of these? To make this concrete, imagine how an employee would feel in four different scenarios: where they get a year-end bonus of either $100 or $10,000, and where this happens after *expecting* a bonus of either $100 or $10,000. Some views suggest that emotions should be more positive when outcomes and/or expectations are more positive, so that, for instance, the employee might be happiest in the case where they expect the larger bonus, and then actually get the larger bonus (e.g., Carver, 2015; Moors et al., 2021; Neville et al., 2021; Smith & Lazarus, 1993). On the other hand, other views suggest that "happiness is reality minus expectations", as put by radio host Tom Magliozzi. More precisely, these views suggest that emotional valence will be driven by the difference between outcomes and expectations, in terms of their reward-value—i.e., by reward prediction errors, or RPEs (e.g., Bennett et al., 2022; Eldar et al., 2016; Emanuel & Eldar, 2023; Loomes & Sugden, 1986). This would mean our employee should feel positive emotions when their bonus is better than expected, negative emotions when it is worse than expected, and neutral when their expectations are met. And still other views suggest that emotional valence is driven by a combination of these factors (e.g., Ding et al., 2025; Rutledge et al., 2014). Yet, empirical work on these ideas has provided mixed results, leaving it unclear what role expectations, outcomes, and mismatches between these play in generating emotions. The current work aims to reconcile these mixed results by showing how seemingly conflicting results could be consistent with RPE-focused theories, after accounting for multiple forms of RPEs and their dynamics over time.

As mentioned, one tradition within both classic and modern emotion theories suggests that emotional valence tracks the value of expected and experienced outcomes. These views are often based on the idea that emotions signal opportunities and threats in the environment—that is, they signal situations that either facilitate or hinder one's goals, needs, and desires. For example, this role for expectations and outcomes can be seen in evolutionary perspectives which suggest that emotions occur "in response to stimuli or situations that are actually, or potentially, rewarding or punishing," and that they "represent the organism's overall experience of reward and punishment" (Mendl et al., 2010, p. 2897). Similarly, Moors' goal-directed theory argues that one "feels happy when she gets what she wants, [and] also when she anticipates getting what she wants" (Moors et al., 2021, p. 149)—in other words, in response to positive outcomes and expectations. Related ideas are present within control-theoretic views and some appraisal theories, which suggest that emotional valence tracks one's progress or rate of progress towards desired end-states—in other words, it should track the value of one's outcomes, defined in terms of how good one's progress or rate of progress is (Carver, 2015; Smith & Lazarus, 1993). All of these theories converge on the prediction that emotional valence should track value, and thus be more positive when outcomes and/or when expectations are more positive. We refer to these perspectives as value-based views.

In contrast, other work has proposed an important role for RPEs in generating emotions, such that people should feel more positive the more outcomes exceed expectations, and feel more negative the more outcomes fall below expectations (e.g., Bennett et al., 2022; Brandstätter & Kriz, 2001; Ding et al., 2025; Eldar et al., 2016; Emanuel & Eldar, 2023; Loomes & Sugden, 1986; Rutledge et al., 2014; Villano et al., 2020; Zeelenberg et al., 2000). Put differently this would manifest as feeling more positive when outcomes are higher, and when expectations are

lower—since this maximizes the RPE, or the difference between outcomes and expectations, as in the case of getting $10,000 when expecting $100.

Yet, within these alternative views, there are different perspectives on how RPEs impact emotions, relative to the outcomes and expectations. Based on empirical results in this area, some views suggest that a combination of RPEs along with outcomes and/or expectations simultaneously drive emotions (e.g., Ding et al., 2025; Rutledge et al., 2014). We refer to views in this category as mixed views. On the other hand, other views take a stronger stance, suggesting that RPEs should drive emotions, while outcomes and expectations should not. For example, some earlier work in this area proposed that RPEs are the primary driver of emotions (Eldar et al., 2016). More recent views suggest that this may be because emotions reflect signals of the perceived need to update one's values—where RPEs are often one such signal, as they track the difference between expected and actual reward. In contrast, outcomes and expectations may instead be linked to other affective[1] states, such as pleasure and pain, or positive and negative evaluations, which directly reflect value representations themselves, including the value of both outcomes and expectations (Bennett et al., 2022; Broekens, 2018; see also Emanuel & Eldar, 2023, for a closely related view). We refer to views of this type as value-updating-based views. Under these views, an employee who receives an unexpectedly large bonus should feel positive emotions to the extent that this RPE is treated as a signal to mentally increase the expected value for future year-end bonuses, while an employee whose bonus perfectly matches their expectations may *like* or *value* their bonus, without necessarily feeling emotions such as happiness. Supporting this value-updating idea, some work suggests that RPEs only drive

[1] We use the term *affect* to refer to the positive-negative dimension of experience (as in Moors et al., 2021), though we note that the current work specifically focuses on explaining the affective valence of *emotional* states, rather than of other states like evaluations.

emotions to the extent that they are informative about changes in value (e.g., when an unexpectedly bad performance is viewed as informative about one's ability vs. reflecting unlearnable randomness; Blain & Rutledge, 2020; see also Emanuel & Eldar, 2023), while recent theorizing suggests that other known drivers of emotions can also be viewed as tracking other value-learning signals (e.g., the relative "advantage" of one action over another; Bennett et al., 2022; see also Emanuel & Eldar, 2023, for related ideas).

Applied to our current questions, these value-updating-based theories thus make two key predictions that distinguish them from other perspectives in this area. First, unlike value-based views, which do not include a role for RPEs in emotion generation, value-updating views suggest that RPEs can and often will generate emotions, with more positive vs. negative emotions tracking more positive vs. negative RPEs. Second, unlike both value-based and mixed views, value-updating-based views predict that outcomes or expectations on their own should *not* directly generate emotions, beyond their role as inputs to computing RPEs (or other value-learning signals).

Yet, despite the growing body of research in this area, the empirical evidence regarding these predictions is inconsistent, and, as a whole, does not clearly fit with any of these three perspectives. Beginning with the first prediction—that RPEs do play a role in emotion generation—existing evidence is currently mixed. Supporting this role of RPEs, many studies do indeed find that positive/negative RPEs trigger corresponding positive/negative emotions. For instance, one study found that students' emotions after receiving exam grades were driven mainly by RPEs in the form of differences between their actual and expected grades (Villano et al., 2020). Similarly, feelings of gratitude have been shown to relate to RPEs in the form of differences between received and expected help from another person (Ding et al., 2025). Similar

results have been observed in several other studies (Bhatia et al., 2019; Brandstätter & Kriz, 2001; McGraw et al., 2005; Mellers et al., 1997, 1999; Neville et al., 2021; Rutledge et al., 2014; Shepperd & McNulty, 2002; Spector, 1956; Vanhasbroeck et al., 2021).

On the other hand, some recent work has failed to find clear or consistent effects of RPEs. For example, the work of Voodla and colleagues (2024) examined emotions during a perceptual decision-making task, where outcomes were manipulated by using easier or harder decision trials, and expectations were manipulated by grouping these trials into easier or harder blocks. Results showed that positive affect in this task was positively related to both outcomes and expectations, with no clear evidence for the role of RPEs. Several other studies show similar results (Blain & Rutledge, 2020; Forbes & Bennett, 2024; Marshall & Brown, 2006; Raz et al., 2024; Vinckier et al., 2018). Furthermore, this lack of RPE effects could potentially be more widespread than currently recognized, since many studies interpreted as showing RPE effects could perhaps reflect the influence of outcomes alone, as these studies did not explicitly test for the distinct contributions of positive outcome effects and negative expectation effects that together constitute an RPE effect (e.g., Blain & Rutledge, 2020; Eldar & Niv, 2015; Keren et al., 2021; Krupić & Corr, 2014; Otto & Eichstaedt, 2018; Rutledge et al., 2015, 2017; Vanhasbroeck et al., 2021; Verinis et al., 1968). Thus, explaining the apparent absence of RPE effects in some cases, or confirming their presence in cases where this is unclear, is something that value-updating accounts still need to address.

Regarding the second prediction—that outcomes and expectations on their own should not directly generate emotions—existing evidence seems to largely contradict this idea. For example, in the previously mentioned study on exam grades, students were slightly happier with more positive grades, even when controlling for the effects of RPEs—suggesting that positive

outcomes also played a more direct, though much smaller, role in generating emotions here (Villano et al., 2020). Similarly, feelings of gratitude were positively impacted by both outcomes (being helped vs. not) and expectations (expecting help), even when controlling for RPEs, suggesting that both outcomes and expectations directly impacted emotions (Ding et al., 2025). These findings are also in line with results from the perceptual decision-making study described previously (Voodla et al., 2024), as well as similar results from other research (Blain & Rutledge, 2020; Forbes & Bennett, 2024; Keren et al., 2021; Neville et al., 2021; Rutledge et al., 2014, 2015, 2017; Vanhasbroeck et al., 2021; Vinckier et al., 2018). Thus, existing evidence on the role of outcomes and expectations appears to suggest that they often drive emotions, seemingly contradicting value-updating-based views.

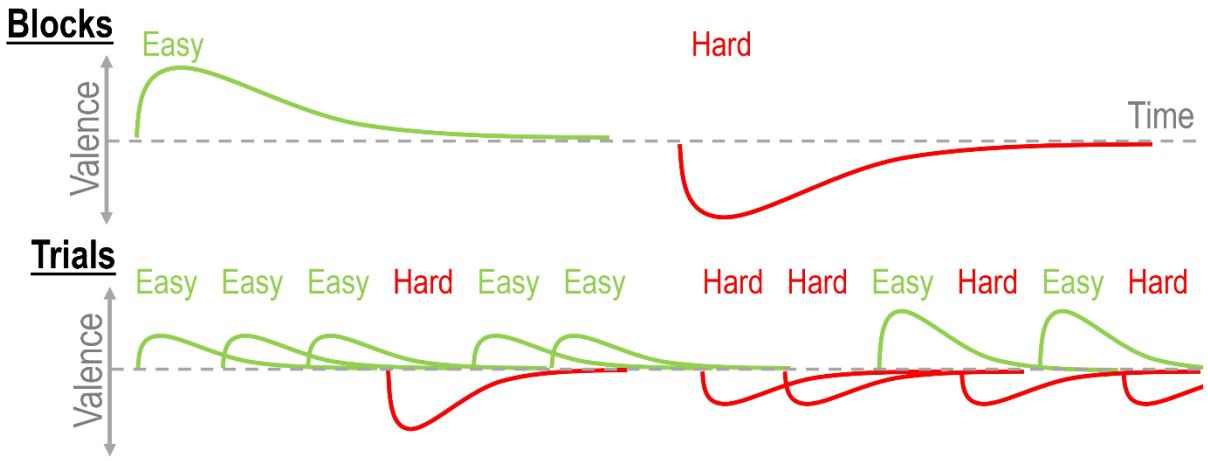
Putting this work together, it appears that outcomes, expectations, and RPEs, can all drive emotional valence in some cases—going against the predictions of a purely value-based or value-updating-based view. Moreover, the extent to which each factor matters appears to differ widely across tasks—for example, with some studies showing little to no clear RPE effect (e.g., Forbes & Bennett, 2024; Voodla et al., 2024) and others showing almost exclusively RPE-driven effects (e.g., Villano et al., 2020). This speaks against mixed views which suggest that all three factors always matter (and current versions of mixed views have not been specified in enough detail to account for the variation in when these factors matter). Thus, at face value, the full set of existing evidence is not adequately explained by any of the three perspectives discussed here. To make progress here requires not just more evidence, but new theoretical insights that can allow for integrating these results within a coherent theoretical framework.

The current work is the first to propose such an integration. Specifically, here we argue that existing findings can be reconciled within a value-updating-based framework, by

considering the effects of multiple overlapping RPEs that may simultaneously drive emotions. To illustrate this idea, consider how these factors could account for the results from the perceptual decision-making study discussed previously (Voodla et al., 2024)—which failed to find RPE effects, and instead found apparent effects of both outcomes and expectations per se. In this study, the researchers looked for evidence of what we will call *trial-level* RPEs—i.e., trials being easier/harder than expected based on the difficulty of the current block—where effects of these RPEs on emotions would manifest as a positive effect of trial-level outcomes (i.e., being happier on easier trials) and a negative effect of trial-level expectations (i.e., being happier in harder blocks). However, we suggest that the effect of these trial-level RPEs may have been obscured by a *second* form of RPE and its residual impact over time: specifically, *block-level* RPEs, due to *blocks* being easier/harder than expected with respect to the *whole task* (see Figure 1 for a conceptual schematic). Such block-level RPEs could account for the apparent *positive* (rather than negative) effect of trial-level expectations observed in this study: that is, perhaps people felt happier on easier blocks not because of their trial-level expectations, but instead because of block-level RPEs, in that people were happy that the current block was easier than expected given the task as a whole. In turn, this block-level RPE effect could have obscured any *negative* effect of trial-level expectations that would be generated by trial-level RPE. Thus, rather than supporting value-based views, which suggest that emotions track both positive outcomes and positive expectations, the results of this study could instead be explained by the overlapping effects of multiple forms of RPEs, consistent with a value-updating based view. This explanation would also be consistent with recent neural and computational work suggesting that multiple types of RPEs can operate in parallel to drive both learning and mood states (Eldar et al., 2018; see Bennett et al., 2022, and Eldar et al., 2016, for related interpretations of past results

Figure 1

Proposed Conceptual Model



Note. We propose that emotional valence in the task used by Voodla et al. (2024) was driven by multiple overlapping RPEs, including both trial-level RPEs (trials being easier or harder than the block average), and block-level RPEs (blocks being easier or harder than the task average). Green vs. red lines illustrate the emotional impact of positive vs. negative RPEs, respectively, that were presumed to occur at different points in the task.

in this area). Moreover, similar arguments could also explain many of the other findings that appear to conflict with value-updating-based views (e.g., Ding et al., 2025; Raz et al., 2024; Rutledge et al., 2017); we return to this point in the General Discussion.

To provide evidence for the current proposal, the present study re-examined the role of RPEs, outcomes, and expectations in a modified version of Voodla et al.'s perceptual decision-making study, to see whether emotion-generation in this task might indeed be driven by these multilevel RPEs. In doing so, we examine three specific questions: 1) whether block-level RPEs contributed to the positive effect of expectations found in the original study, 2) whether trial-level RPEs contributed to emotions in this task after accounting for other overlapping RPE effects, and 3) whether the apparent effects of outcomes and/or expectations per se (beyond their use in computing RPEs) could be explained away when controlling for these multiple forms of

RPEs and their residual effects over time. By accounting for the multiple ways RPEs can drive emotions, the current study aims to provide one of the strongest tests yet of the role of outcomes, expectations, and RPEs in generating emotions. In doing so, this work aims to demonstrate how apparently conflicting results in this area may be unified within a value-updating perspective, with the ultimate aim of shedding light on the fundamental computations underlying the generation of emotions.

Method

The key methods of the current study were based closely off Voodla et al.'s (2024) study. As in that previous study, the current work used a perceptual decision-making task (deciding on the overall motion direction of a cloud of moving dots, left or right), where trial-level *outcomes* were manipulated by using easier or harder trials (with more or less coherent motion—i.e., more or less of the dots moving in the same direction), and where trial-level *expectations* were manipulated by grouping these trials into blocks of primarily easier or harder trials.

While the core logic of this task is the same as in Voodla et al.'s work, several key changes were made to allow for distinguishing effects of trial-level and block-level RPEs. One of the most important changes was using much longer blocks. Since the effects of block-level RPEs on emotions should be strongest at the start of a block, but fade away over time, using longer blocks allows us to isolate effects of trial-level RPEs by looking at the end of the block, as well as to find a key signature of block-level RPEs through examining temporal dynamics across a block. Below, we describe the current task in detail, followed by further details on how we conceptualize this task, as well as the specific changes from Voodla et al.'s design and the rationale for these changes.

Participants

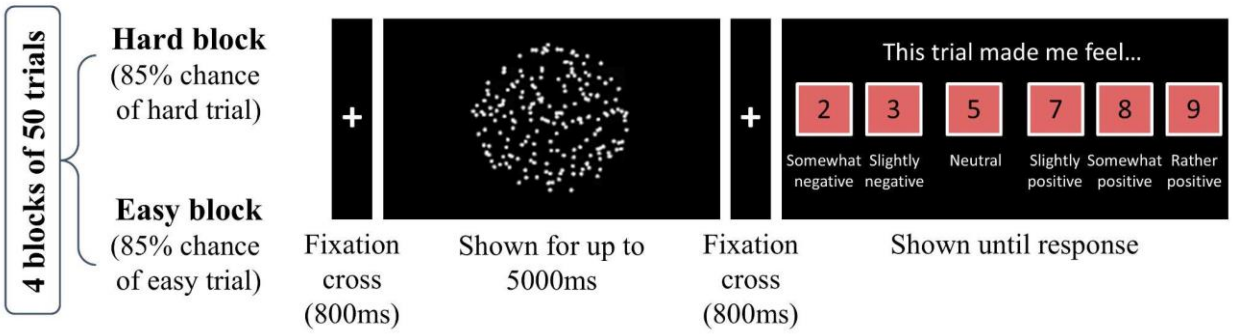
47 participants were recruited from the undergraduate participant pool at KU Leuven in Belgium in exchange for course credit. This sample size was based on practical reasons (the number of participants that could be recruited by the end of the semester). Participants were excluded if they had less than 60% accuracy on trials of the easiest difficulty level. The final sample included 41 participants (M_{age} = 19.78, SD_{age} = 2.15; 30 females, 11 males).

Procedure

The study was conducted online using the Cognition.run platform. The main part of the task involved completing a perceptual decision-making task very similar to the one in Voodla et al.'s original study (see Figure 2). On each trial, participants were shown a cloud of moving dots and were given 5 seconds to select which direction more of the dots were moving in (V key = left, B key = right). The initial task instructions asked participants to respond as quickly and accurately as possible to these decisions. Trials varied across seven levels of difficulty, with coherence levels ranging from 0.1 (hardest) to 0.4 (easiest) in increments of 0.05, where coherence reflected the proportion of dots moving in the same direction (left or right). Following the direction-of-motion judgment, participants rated their current affective state. Specifically, they responded to the question: "This trial made me feel…" (responses made using the following keyboard keys: 2 = "Somewhat negative", 3 = "Slightly negative", 5 = "Neutral", 7 = "Slightly positive", 8 = "Somewhat positive", 9 = "Rather positive"). In terms of timing, each trial involved displaying a fixation cross for 800ms, followed by the dot cloud (shown until the participant responded, or for a maximum of 5000ms), a 500-ms pause, and then the affect rating scale (shown until the participant responded). If no response was made for the motion direction judgment, a message saying 'Too late!' appeared, and no affect rating was made.

Figure 2

Task Structure



The main part of the task consisted of 200 trials, structured into four blocks of 50 trials each. There were two types of blocks: easy and hard. The difficulty levels for the trials within a block were randomly selected based on the following probability distribution: there was an 85% chance of selecting trials at the easiest difficulty level in the easy block, or at the hardest difficulty level in the hard block, with the remaining probability distributed evenly across the other six difficulty levels. The four blocks were presented in a fixed order (easy, hard, easy, hard).

After every 25 trials, at the middle and endpoint of each block, participants were given a short break. During this break, participants rated how difficult they found the previous section (on a 100-point sliding scale ranging from “Very easy” to “Very difficult”), which served as a manipulation check, and they rated the perceived accuracy of their affect judgments in the previous section. They were then given feedback on their performance in that section (the percentage of correct responses, and average response time), as well as a brief reminder of the task instructions.

The structure of the study as a whole was as follows. Participants started by going through instructions for the perceptual decision-making task and several rounds of training trials.

Participants were then introduced to the affect scale and given practice using it. The instructions for the affect scale were designed to encourage participants to carefully rate their emotional state on each trial, and not simply give the same answer to all trials, or rely on some other non-emotional basis for their judgments, such as their decision confidence. After these instructions, participants completed the 200 trials of the main task, followed by a series of questions about their experience in the task and a demographic questionnaire.

Conceptualization of Reward Computations in this Task

To fully understand the logic of the current study, it is important to clarify the presumed nature of the reward signals in this task. We assume that participants' primary goal in this task is to be accurate, thus, accuracy on each trial can be viewed as the objective, external reward in this task, while expected accuracy can be viewed as the reward expectation. However, since participants were not provided with objective performance feedback after each trial, we assume they are *estimating* their accuracy on each trial, based on information such as the coherence of the dot motion and their resulting decision confidence (e.g., if the dots were clearly observed to be moving in one direction, so that the participant felt confident that they made an accurate decision, vs. if they did not perceive a clear direction of motion, and thus subjectively felt like they were guessing and had less confidence in their decision). Given that rewards must be estimated mentally in order for them to impact people's emotions, these *internal* reward signals may thus be best characterized in terms of decision confidence—participants' subjective estimate of their accuracy in the task (mirroring Guggenmos et al., 2016). Indeed, supporting this idea, past work shows that people's decision confidence reliably tracks both decision accuracy and motion coherence (trial difficulty) in this task (Voodla et al., 2025; Zylberberg et al., 2012). Moreover, confidence has been linked to reward processing and mesolimbic reinforcement

learning signals in perceptual learning tasks (Guggenmos et al., 2016), as well as across a variety of domains (Sharot et al., 2023).

In light of this, the key manipulations in the task can be interpreted as follows. Easier vs. harder *trials* can be viewed as manipulating both subjective and objective trial-level outcomes, by increasing vs. decreasing both estimated and actual accuracy on that trial. Easier vs. harder *blocks* can be viewed as manipulating trial-level expectations by affecting one's expected accuracy on the upcoming trial, based on one's experience with previous trials in that block.

In terms of the mental RPE computations thought to drive emotions, trial-level RPEs can then be defined as the difference between the estimated outcome of a given trial, compared to one's expectations for that trial, based on the current block, while block-level RPEs can be defined as the difference between the estimated accuracy on the current *block,* compared to expectations for that block, based on one's experience with the task as a whole.

Differences from Voodla et al.'s Original Study

The current study made several changes to Voodla et al.'s study design to help find effects of trial-level RPEs, and to distinguish these from the effects of block-level RPEs. As mentioned, one of the primary changes involved extending the duration of each block—from six trials to 50 trials, before changing the block's difficulty level. If block-level RPEs are triggered near the start of a block when people first realize the block is easier/harder than expected (relative to their task-based expectations), longer blocks give more time for the emotions created by these initial block-level RPEs to fade away, as block-level expectations get updated. Evidence that emotions follow this pattern would provide evidence for a role of block-level RPEs in this task. In addition, this change should allow for observing effects of trial-level RPEs when looking at the *end* of a block (including finding the negative effects of trial-level expectations, which

should lead to feeling happier on *harder* blocks), even if block-level RPEs (and the corresponding tendency to feel happier on *easier* blocks) dominate early in the block.

In addition to extending the blocks, another important change was making more of the trials within a block match the block-based expectations. That is, we increased the percentage of trials that were at the easiest difficulty level in the easy blocks, or at hardest difficulty level in the hard blocks, from 70% to 85%. This change was designed to further facilitate the updating of block-level expectations and thus the fading of emotions associated with block-level RPEs over time. In addition, by forming stronger expectations of a given block's difficulty and associated performance accuracy, this should also allow for more clearly observing effects of trial-level RPEs, since expectations about the current block must exist in order to be compared to outcomes on a given trial.

Several other changes were made to further increase the chance of observing trial-level RPE effects. For instance, Voodla et al.'s study included labels before each block and trial that alerted participants to the current block's difficulty level. Here, we removed these labels, as they may have exacerbated the salience, and therefore impact, of block-level RPEs on emotional responses, thus further obscuring the effects of trial-level RPEs. These labels were included in the original task to make sure participants were aware of the type of block they were in. However, anecdotally, in the current version of the task, the difference between blocks was obvious enough that the block type could be easily identified within a few trials, even without these labels. Thus, removing these labels could provide a cleaner test of the role of trial-level RPEs, especially when focusing later in a block, when these expectations should be well-established and thus the labels made unnecessary.

In addition, we wanted to ensure that trial-level RPEs were computed relative to expectations based on the current *block's* difficulty level—the form of trial-level RPE that the current analyses test for—rather than based on the current *trial's* difficulty level. This latter form of expectation could have been used, for example, if participants realized they were in an easy trial within the first several milliseconds, and then used their expectation for what easy trials are like in this task when computing RPEs. Such a computation would presumably not produce many RPE signals, since, for example, a given easy trial, once identified as such, would typically match expectations for what easy trials are like in the task. To try to avoid this situation, rather than using two clearly distinct coherence levels (easy vs. hard trials), as in Voodla et al.'s original study, we used seven coherence levels of gradually varying difficulty. These more subtle gradations should reduce the tendency to rapidly categorize the trials based on their perceived difficulty, and thus encourage the trials in a given block to be categorized similarly and have a similar expectation applied across them.

We also wanted to help ensure that participants' affect judgments were in fact based on their emotions, rather than something else, such as evaluations of one's perceived accuracy on a given trial (i.e., of one's decision confidence). As discussed, value-updating-based perspectives suggest that emotions should track RPEs or other related computations, while evaluations should track values themselves. Thus, if affect judgments were merely based on evaluations in Voodla et al.'s original study, it could provide another explanation for why these judgments primarily depended on outcomes (with what could be considered a small biasing effect of expectations), and why there was no clear evidence of RPE effects. We therefore added further instructions to the current task which emphasized that the affect ratings should be based on emotions, and not necessarily on evaluations of one's confidence. In addition, we introduced a "Neutral" option in

the affect rating scale, which was not included in the original study, to reduce the chance of participants turning to other factors as a basis for their response in cases where no emotion was felt. Because adding a response option would have made hand placements on the keyboard uncomfortable, we chose to remove the most extreme negative response option from Voodla et al.'s scale, since this option was only rarely used in the original study. In addition, we modified the affect scale by using milder labels for the most extreme responses (e.g., "Rather positive" instead of "Quite positive"), given the generally mild nature of emotions in this task.

Finally, we modified the breaks between sections, where objective performance feedback was given. Specifically, including this performance feedback twice per block, rather than once per block, allowed for uncorrelating this feedback (which should track the previous section's difficulty) from the difficulty of the subsequent section. This allowed for statistically controlling for emotions generated by this feedback and separating them from the effects of the block difficulty. The ratings made during these breaks were also newly added: difficulty ratings for the previous block were included as a manipulation check, and ratings of the accuracy of one's affect judgments were added to encourage participants to continue making accurate judgments throughout the task.

Data Preparation and Coding

Trials were excluded if reaction times for the motion-direction decision were below 200ms or above 4000ms, or if participants took over 20 seconds to respond to the affect judgment on that trial. This led to 1.06% of trials being excluded.

Data were analyzed using R v. 4.4.2 (R Core Team, 2024), and the following packages: lmerTest v. 3.1.3 (Kuznetsova et al., 2017) for multilevel regressions, rms v. 7.0-0 (Harrell Jr, 2025) for modelling restricted cubic splines, emmeans v. 1.10.7 (Lenth, 2019) for estimating and

comparing marginal effects within a given model (e.g., estimating the effect of one variable at a given level of another) including in models involving restricted cubic splines, and multcomp v. 1.4.28 (Hothorn et al., 2008) for general linear hypothesis testing. All analyses were performed as multilevel regressions, or as multilevel logistic regressions for binary outcome variables, with random intercepts for each participant.

For all analyses, the following variable coding was used. Block type was effect coded (1 = easy, -1 = hard). Trial coherence, which reflected how easy or difficult each trial was (i.e., the proportion of dots moving in the same direction), originally ranged from 0.1 (hardest) to 0.4 (easiest) and was recentered at its mean value of 0.25. Average block coherence was conceptually equivalent to the block type variable, but was coded in terms of units of coherence so it was on the same scale as the trial coherence variable: specifically, it was coded as the average coherence level for the current block type (i.e., 0.37 for easy blocks, 0.12 for hard blocks), and was then recentered at its mean value of 0.25. Accuracy on a given trial was effect-coded (1 = correct, -1 = incorrect) when used as a predictor, or else coded as 1 = correct and 0 = incorrect when used as an outcome. Trial number within a block was coded so that it ranged from -49 for the first trial in a block to 0 for the last, to allow main effects for the variables it interacted with to be estimated at the end of a block. Performance feedback was coded as the proportion of correct trials in the previous set of 25 trials (or, for the first 25 trials in the task, the proportion correct in the last training block) and ranged from 0 to 1. The number of trials since the last feedback ranged from 1 to 25, and reflected the trial number starting from the last time performance feedback was given. Trial number within experiment reflected the trial number starting at the first trial after the training was completed. Six different variables were computed to capture previous trials' affect ratings: these reflected affect ratings from the previous trial,

from two trials back, etc., all the way up to 6 trials back, excluding trials from the start of the experiment where the relevant previous ratings did not exist.

Transparency and Openness

The current paper reports how the sample size was determined, all data exclusions, and all manipulations and measures in the study. This study was preregistered after data collection was complete but before looking at the data. The preregistration specified the hypothesis regarding our second question (the effect of trial-level RPEs in this task) and specified the design, analysis, and exclusion criteria for this study. Additional non-preregistered analyses are labeled as exploratory. The preregistration, data, analysis code, and study materials are available at https://osf.io/v5ubg/?view_only=bf3e05ef57f74e988a9941892caadf18.

Results

Preliminary Analyses

Before turning to our main questions, preliminary manipulation checks tested whether harder blocks were in fact harder for participants than easier blocks. A multilevel regression confirmed that, compared to easier blocks, harder blocks showed lower accuracy (proportion correct: hard blocks: $M = 0.66$, $SD = 0.47$; easy blocks: $M = 0.84$, $SD = 0.36$; difference: OR = 1.69, 95% CI [1.60, 1.78], $p < .001$) and higher difficulty ratings (hard blocks: $M = 75.14$, $SD = 17.90$; easy blocks: $M = 53.37$, $SD = 23.02$; $\beta = -0.46$, 95% CI [-0.54, -0.39], $p < .001$). This confirms that block difficulty was successfully manipulated.

Block-Level RPEs Impact Affective Valence

Our first analyses examined whether the apparent positive effect of trial-level expectations found in Voodla et al.'s work—i.e., the positive effect of block type (feeling more positive on easier blocks)—could have instead been driven by block-level RPEs, due to starting a

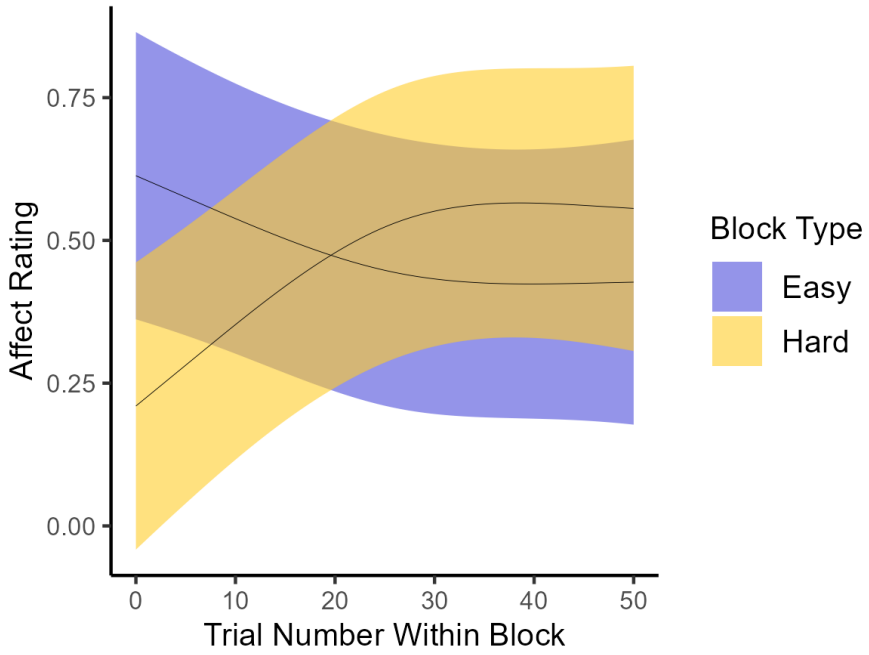
block that is easier or harder than expected relative to the task-average. If this was the case, in the current task with its longer blocks, we should observe a temporary increase in positive affect at the start of an easier block (or decrease in positive affect at the start of a harder block) as the block unexpectedly changes in difficulty, and this should diminish over time as people update their representations of the task's difficulty level, and the effects of block-level RPEs fade away.

To test this, we fit an exploratory model where affect was predicted by block type, trial number within a block, and trial coherence, with block type allowed to interact with trial number and coherence. Trial number was modeled using restricted cubic splines, to capture potentially nonlinear changes over time in the effect of block type. Including coherence in the analysis allowed for estimating block type effects after controlling for the difficulty of each trial, so that these effects capture the different affective experience of doing a given difficulty-level trial within an easy block vs. doing that same trial within a hard block.

This analysis revealed that, at the start of a block, participants indeed felt more positive affect on easier vs. harder blocks (b = 0.20; β = 0.15, 95% CI [0.09, 0.21]; $p < .001$), even after controlling for trial difficulty, but this effect decreased across the block, and eventually became non-significant by the end of the block (b = -0.06; β = -0.04, 95% CI [-0.11, 0.01]; $p = .11$; interaction of block difficulty and trial number: $F(2, 8065) = 21.54$, $p < .001$); see Figure 3. This initial positive effect of block type that fades over time is consistent with the idea that block-level RPEs contributed to affect in this task. It also suggests that this positive block type effect should not be attributed to trial-level expectations per se, as any effects of trial-level expectations should presumably be consistent or even strengthen across a block, rather than fading away. Linking this result to Voodla et al.'s work, it suggests that their original finding, where people felt more positive on easier blocks, was likely due to this initial effect of block-level RPEs on

Figure 3

Effect of Block Difficulty on Affect Across the Block



Note. At the start of a new block, affect was initially more positive in easier blocks than harder blocks (controlling for trial difficulty), but this effect diminished over time across the block, consistent with this initial effect being driven by block-level RPEs. Error bands indicate 95% confidence intervals.

affect, with this effect predominating in their task due to the use of much shorter blocks (6 vs. 50 trials), which did not allow time for it to fade away.[2]

Trial-Level RPEs Impact Affective Valence

Our second question was whether we could also find effects of trial-level RPEs in this

[2] Additional analyses confirmed that we replicate Voodla et al.'s results using their analysis (predicting affect from block type, trial coherence, and their interaction) when only analyzing the first 6 trials of each block in our task, but not when including all trials in the block. In particular, an exploratory analysis on the first 6 trials showed a positive effect of trial coherence (b = 1.75; β = 0.17, 95% CI [0.08, 0.29]; p < .001), as well as of block type (b = 0.23; β = 0.17, 95% CI [0.07, 0.28]; p = .001), while a preregistered preliminary analysis using all trials found a positive effect of trial coherence (b = 2.06; β = 0.21, 95% CI [0.18, 0.25]; p < .001), and no significant effect of block type after controlling for trial coherence (b = 0.00; β = 0.00, 95% CI [-0.03, 0.04]; p = .96), which makes sense given the changes in the block type effect over time shown in the main text.

task, at least towards the end of a block, after effects of block-level RPEs have faded away. This trial-level RPE effect should manifest as the combination of a positive effect of trial-level outcomes (as found by Voodla et al.), and a negative effect of trial-level expectations (not seen in Voodla et al.'s work). Preliminary indications of this latter effect can be seen in the previous analysis. Specifically, inspecting Figure 3 suggests that the impact of block-level RPE had largely faded by trial 30 to 40 (as indicated by the leveling off of the curves for the different block types), and the effect of block type seems to reverse at that point. Though this reverse effect was not significant in the previous analysis, its direction—i.e., feeling more positive on *harder* blocks (after controlling for trial difficulty)—is consistent with a negative effect of trial-level expectations, as part of trial-level RPEs effects. These initial results suggest that we may indeed find clear evidence of trial-level RPE effects, if we focus on results towards the end of a block, while also using a more carefully controlled analysis.

The main preregistered analysis thus tested for these trial-level RPE effects, by estimating effects at the end of each block, while also controlling for several additional factors that might have obscured the effect of trial-level RPEs. To do so, several control variables were added to the previous analysis: objective accuracy on each trial (another proxy for subjective trial outcomes, a component of mental RPE computations); the previously seen performance feedback, the number of trials since that performance feedback was seen, and the interaction of these two variables (capturing any feedback-related RPE effects that may fade over time); the number of trials since the start of the experiment (capturing any RPEs from the task as a whole becoming, say, unexpectedly easier over time due to perceptual learning); and affect ratings from the previous six trials (capturing effects of any RPEs from earlier in the task, including those from trial-level RPEs on earlier trials). To estimate block type effects on the same scale as the

trial coherence effect, the block type variable was recoded as average block coherence (i.e., average trial coherence for that block type). As per the preregistration, trial number in the block was modelled as a linear effect here for simplicity.[3] In addition, by centering this variable with the zero value at the last trial in a block, the main effect of block type (i.e., average block coherence) was estimated at the end of the block, after effects of block-level RPEs should have faded.

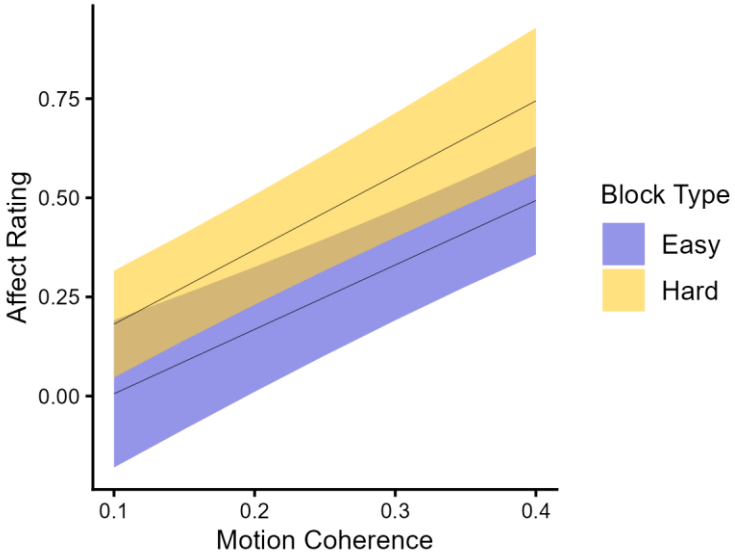
If trial-level RPEs were indeed driving affect here, then this analysis should show two key effects: 1) a positive effect of trial-level outcomes, which should manifest as a positive effect of trial coherence—in other words, feeling more positive on easier trials—and, 2) a negative effect of trial-level expectations, which should manifest as a negative effect of block type (i.e., average block coherence)—in other words, feeling more positive on harder blocks, after controlling for trial difficulty. The full model indeed revealed these two effects. Specifically, as shown in Figure 4, there was both a positive effect of trial coherence (b = 1.75; β = 0.18, 95% CI [0.15, 0.22]; $p < .001$), and a significant negative effect of block type (i.e., average block coherence) after controlling for trial coherence (b = -1.34; β = -0.12, 95% CI [-0.17, -0.07]; $p <$ .001). This pair of results is consistent with the idea that trial-level RPEs do contribute to emotional responses in this task (see Appendix for further support of this interpretation). The fact that we only observed this pattern when looking at the end of a block once block-level RPE effects had faded, and after statistically accounting for various forms of RPEs and their impacts

[3] An additional preregistered analysis was run where effects of block type and feedback over time were modelled using restricted cubic splines to account for potential non-linearity in how these effects change over time. The key results reported here replicated in this version of the analysis.

over time, also suggests that trial-level RPEs may indeed have contributed to affect in Voodla et al.'s original study, but that this effect was likely obscured by these other factors.[4]

Figure 4

Effects of Trial Difficulty and Block Difficulty on Affect



Note. When looking at the end of a block, participants reported more positive affect when there were more positive trial-level outcomes—i.e., on easier, higher coherence, trials—and when there were more negative trial-level expectations—i.e., on harder blocks, after controlling trial difficulty—consistent with the idea that trial-level RPEs can drive affect in this task. Displayed results reflect estimates at the end of a block (i.e., after block-level RPE effects should have faded), and after controlling for several other variables to capture other RPEs and their effects over time (see main text). Error bands indicate 95% confidence intervals.

[4] Although not the primary purpose of this analysis, additional results from the main analysis shed some light on which of these other factors may have been more important to account for. In particular, aside from the results reported in the main text, and the interaction of block type (i.e., average block coherence) with trial number since start of block (mirroring results from the previous analysis), there was a significant effect of accuracy on the current trial ($b = 0.45$; $\beta = 0.14$, 95% CI [0.12, 0.16]; $p < .001$). Since objective accuracy was not strictly known to participants, but instead could only be estimated based on their subjective confidence, this factor likely serves as another proxy for the effect of estimated trial-level outcomes (i.e., confidence), which may have driven affect as part of trial-level RPE effects. The only other significant effects were a small positive effect of trial number within the experiment ($b = 0.0008$; $\beta = 0.03$, 95% CI [0.01, 0.05], $p < .001$), suggesting that participants got slightly happier as the experiment progressed (perhaps due to perceptual learning leading to task performance becoming better than expected over time), and positive effects of affect ratings on each of the previous six trials (bs ranging from 0.21 to 0.03, βs ranging from 0.21 to 0.03, ps ranging from < .001 to .005), which may have accounted for various forms of RPEs from earlier in the task (including block-level RPEs, or trial-level RPEs on previous trials, which are not otherwise accounted for in the analysis).

Outcomes And Expectations Do Not Impact Affective Valence, Beyond their Role in RPEs

While the two key effects in the previous analysis provide evidence of trial-level RPEs driving affect, it is still possible that there is some additional contribution of trial-level outcomes or expectations per se, beyond their role in computing RPEs—e.g., so that, hypothetically, people might feel more positive on easier trials or easier blocks, even in cases where there was no RPE effect at play. One way to look for such contributions would be to compare the magnitude of the outcome and expectation effects in the previous analysis. In particular, if these effects were purely driven by trial-level RPEs, then, since RPE = outcome - expectation, we should see a positive effect of outcomes that is equal in magnitude to the negative effect of expectation, when both effects are on the same scale. In contrast, an asymmetrical effect, where the positive effect of outcomes is larger than the negative effect of expectations, would indicate some additional contribution of outcomes and/or expectations per se—since, for example, any additional positive impact of outcomes per se would increase the size of the positive outcome effect in the regression results, while an additional positive impact of expectations per se would decrease the size of the negative expectation effect in the regression results (i.e., make it less strongly negative). Thus, this approach can test whether either of these factors have additional contributions, beyond their role in computing RPEs. To test for such asymmetries, the magnitude of the outcome and expectation effects from the previous analysis were compared using linear hypothesis testing (i.e., computing $b_{\text{outcome}} + b_{\text{expectation}}$) in an exploratory follow up test. Results showed that there was no significant asymmetry here (b = 0.41, 95% CI [-0.04, 0.86], p = .07). Thus, there is no clear evidence that affect was directly influenced by outcomes or expectations in this task, beyond their role in computing RPEs—in contrast to Voodla et al.'s original conclusions.

Moreover, while these additional contributions were not statistically significant, we can also compare the size of their estimated contribution to that of trial-level RPEs. In particular, the estimate of this asymmetry computed above ($b_{outcome} + b_{expectation}$) also represents the total additional contribution of outcomes and expectations per se, beyond their role in RPEs; see Appendix for derivation. And, while it is mathematically impossible to compute the exact impact of RPEs in this task, if any contributions of outcomes or expectations per se are assumed to be non-negative, then we can compute bounds on the possible impact of RPEs: i.e., it should be between the magnitude of $b_{expectation}$ and $b_{outcome}$, or between 1.34 and 1.75 here (95% CIs: [0.78, 1.90], [1.42, 2.08]); again, see Appendix for derivation. Thus, comparing the estimates and confidence intervals for these different values suggests that even if there was a real contribution of outcomes and/or expectations per se that merely did not reach significance in this study, it would have to be much smaller than that of trial-level RPEs, further supporting the primacy of RPEs in this task.

Discussion

The current study re-examined the role of outcomes, expectations, and reward prediction errors (RPEs) in generating emotions. Existing theories diverge on which of these factors should affect emotional valence, with value-based views suggesting a role for outcomes and expectations (but not RPEs), value-updating views suggesting a role for RPEs (but not expectations or outcomes per se), and mixed views suggesting that these factors simultaneously matter. Yet empirical work has not been fully consistent with any of these views. The current work proposed that we can reconcile these mixed results within a coherent theoretical framework—specifically, within a value-updating framework—if we consider the impacts of multiple forms of RPEs and their dynamics over time. Here, we showed how this idea can

explain one piece of apparent counter-evidence from previous work: a study by Voodla et al. (2024) which suggested that affect in a perceptual decision-making task was driven by outcomes and expectations per se, and not by RPEs. In contrast, through carefully distinguishing and controlling for different types of RPEs, we showed that affect in this task was instead driven by the overlapping effects of multilevel RPEs. Specifically, we showed that two distinct forms of RPEs drive affect in this task: trial-level RPEs (trials being easier/harder than expected given the current block of trials), and block-level RPEs (blocks being easier/harder than expected given the task average), with no clear effects of outcomes or expectations per se, beyond their role in these RPE computations.

Though we focused on explaining one particular set of past findings, similar explanations could also account for many of the other findings that appear to conflict with predictions of a value-updating-based view (e.g., Blain & Rutledge, 2020; Ding et al., 2025; Forbes & Bennett, 2024; Raz et al., 2024; Rutledge et al., 2014; Vanhasbroeck et al., 2021; Villano et al., 2020). For example, in some studies (e.g., Ding et al., 2025; Rutledge et al., 2014), apparent effects of outcomes or expectations per se could be driven by additional forms of RPEs not considered in the original work (e.g., outcomes or trial-specific expectations differing from expectations based on the task average—similar to the block-level RPEs considered here—rather than outcomes on a given trial differing from trial-specific expectations—similar to the trial-level RPEs considered here). Even in cases when there is only one type of RPE at play, apparent effects of outcomes or expectations per se could also reflect residual emotions from RPEs earlier in the task, if these are not statistically controlled for (as in, e.g., Ding et al., 2025; Voodla et al., 2024). Thus, considering the impacts of multiple forms of RPEs and their dynamics over time could also allow for integrating these other conflicting findings into a value-updating framework. The

current results increase the plausibility of this idea, and highlight the need to test whether other findings can indeed be explained in this way.

More broadly, more fully accounting for the effects of RPEs may also provide other ways to integrate existing counter-evidence within a value-updating perspective. For example, while many studies use objective manipulations of outcomes and expectations (and thus RPEs), these may fail to fully reflect the *subjective* RPE representations that presumably drive emotions—e.g., if people's expectations are not affected, or are only somewhat affected, by objective manipulations (as suggested by Vinckier et al., 2018, to explain their lack of RPE effect), or if people show biases in subjective value, such as interpreting outcomes more positively when in a positive mood (Eldar & Niv, 2015; Vinckier et al., 2018). Indeed, in the current work, these factors could perhaps have contributed to the slight, non-significant effect of outcome or expectations per se, that was suggested by our final analysis. On the other hand, self-reported expectations or outcomes (as used in, e.g., Villano et al., 2020) may risk introducing noise into estimates of RPEs, and may also contain biases that can produce spurious results (see Marshall & Brown, 2006). In either case, failing to fully control for the effects of mental RPE representations could lead to apparent effects of outcomes and expectations per se, beyond their role in RPEs (see Buttrick et al., 2020; Marshall & Brown, 2006). In addition, some of the apparent counter-evidence in this area could be due to how affect is measured, if the measures captured *evaluations*, in addition to, or instead of, emotions—since, according to value-updating views, evaluations may be driven by outcomes and expectations per se, with emotions driven by RPEs. While the current work explicitly asked participants to rate emotions (and not evaluations of their confidence), past work has not typically clarified this distinction to participants, despite its theoretical importance. Future work should thus continue to explore the best ways to capture

mental RPE representations and their impacts on emotions, and examine whether this allows for further reconciling past findings within a value-updating-based perspective.

The current research also provides other important guidance for future work looking at the role of outcomes, expectations, and RPEs in emotion-generation. One major takeaway is that future work should make sure to fully consider all forms of RPEs that could occur in a given task, and the temporal dynamics of these effects—especially in cases where affect from other forms of RPEs or from RPEs earlier in the task could be correlated with variables of interest on a given trial (as is common in studies where trials are grouped into blocks of similar value, e.g. Ding et al., 2025). Moreover, future work would benefit from accounting for the temporal dynamics of affect in more detail, even when they deviate from some pre-specified form (e.g., an exponential decay of objective task variables, as in Rutledge et al., 2014). The current work accomplished this by having participants rate their affect on every trial, and controlling for affect ratings from several previous trials (vs. having ratings only every few trials, and assuming trials in between follow a specific decay function, as is common in past work; e.g., Blain & Rutledge, 2020; Forbes & Bennett, 2024; Rutledge et al., 2014). The current approach allowed for more flexibly capturing residual affect from earlier in the task, including from unmeasured sources of affect, as well as from trial-to-trial, person-to-person, and source-to-source variability in the impact or decay of any given source of affect. In addition, when looking for evidence of RPEs, it is important to confirm that *both* expectations and outcomes play their presumed role—as was done here for trial-level RPEs—to ensure that an apparent RPE effect is not, say, driven purely by outcomes (see Marshall & Brown, 2006).[5] Finally, it is important to be mindful of the

[5] In the current task, this could not be checked directly for block-level RPEs, as this would require including different versions of the task that varied the difficulty of the task as a whole. However, the temporal dynamics of the block-level RPE effects observed here provide additional support for our interpretation of them. In particular, the observed dynamics—i.e., where people were initially happier in easier blocks, but this decayed over

intrinsic relationship between outcomes, expectations, and RPEs (i.e., the fact that RPE = outcome – expectation), which often makes it impossible to statistically estimate all three of these effects at the same time. The current work addressed this by not dissociating the effects of expectations vs. outcomes per se, and instead just separating these from the effects of RPEs. (For attempts to further dissociate these three factors, see: Ding et al., 2025; Neville et al., 2021; Rutledge et al., 2014; Villano et al., 2020). However, future work and interpretations of past work should pay careful attention to this issue, as the estimates and interpretations of these effects will depend on which of these variables one chooses to model (see Appendix for guidance on these interpretations, and also Forbes & Bennett, 2024).

One question raised by the current results is how to best interpret the effects of RPEs in this task. We specifically interpreted these RPE effects as driven by the difference between estimated vs. expected performance accuracy (with estimated performance accuracy being the presumed reward signal here). Yet these effects could also be interpreted within a predictive processing framework (as in Voodla et al., 2024), as driven by the difference between the actual and expected *ease of perceptual processing* (more precisely, by differences in the actual vs. expected rate of resolving the perceptual prediction errors used to identify the dots' motion; see Clark, 2013; Van de Cruys, 2017; Voodla et al., 2024). One way to test between these possibilities would be to use a much easier version of the current task, to see whether emotions are still impacted by variations in motion coherence (i.e., changes in perceptual ease of processing), even when performance is at 100% accuracy in all cases. More broadly, even if emotions in this task are driven by unexpected ease of processing, this could still be viewed as a

time—are consistent with an RPE effect, since RPEs should fade away as expectations are updated, but they are inconsistent with an effect of block-level outcomes per se (i.e., being happier because the current block was easy and led to high accuracy) since this type of effect should presumably persist or strengthen across the block. Thus, looking at the temporal dynamics of potential RPE effects may also provide another clue to their origins.

form of reward prediction error, given the widespread and well-supported assumption that sense-making is intrinsically rewarding, while processing difficulty and mental effort have an intrinsic cost (Alter & Oppenheimer, 2009; Chater & Loewenstein, 2016; Gopnik, 2000; Kurzban, 2016; Schmidhuber, 2008; Topolinski & Reber, 2010). Thus, either interpretation of the current results would be consistent with the idea that reward prediction errors drive emotions.

More broadly, the current work provides some of the strongest evidence yet for a value-updating perspective on emotions—by supporting a key prediction of this view that differs from that of value-based or mixed views, and by providing a way to reconcile seemingly inconsistent evidence within a value-updating framework. This value-updating perspective in turn has further implications for understanding the role of RPEs in the generation of emotions, and the generalizability of the current results. For instance, this perspective suggests one important moderator of the link between RPE and emotions: the perceived relevance of these RPEs for value-based learning. Indeed, a recent empirical study supports the idea that RPEs no longer drive emotions in cases where they are not relevant for learning (e.g., when RPEs are based on the random, unlearnable amount of points associated with correct guesses on each trial, rather than on learnable outcomes that help make better predictions; Blain & Rutledge, 2020).

Interestingly, applied to the current task, this suggests that RPEs *were* considered relevant for value-learning here, despite value-learning not being strictly necessary for the task—perhaps because people were learning about the value of continuing to engage in the task, or the value of their current decision strategy. Supporting this possibility, recent work suggests that this type of learning-signal indeed may be spontaneously computed and drive emotions, even when not strictly required by the task (Keren et al., 2021). In particular, recent work has re-examined a commonly studied gambling task that also does not require value-learning (e.g., as used by

Rutledge et al., 2014), and suggested that the RPEs that drive affect in this task may indeed be learning-relevant in this way: since these RPE effects were better modeled as comparisons between gamble outcomes and the expected average reward per trial—exactly the form of RPE that could help learn about the value of the task as a whole, or of the current decision strategy—rather than as comparisons between gamble outcomes and the gamble's mathematical expected value, as initially assumed—a form of RPE that is irrelevant to any form of value-learning, since each gamble is only encountered once (for related interpretations of these gambling tasks, see Bennett et al., 2022; Eldar et al., 2016). Future work could more directly test the learning-relevance of RPEs in such cases (e.g., in these gambling tasks, as well as the current task) by seeing whether the RPEs that drive emotions in these tasks also show corresponding effects on value-learning—e.g., by giving people a free choice to continue with the same task or try a new task, or by adding a reinforcement-learning component directly to the task (as in Ding et al., 2025 or Forbes & Bennett, 2024). More generally, further establishing more direct links to value-learning, including the moderating role of learning-relevance, will be an important future direction for value-updating perspectives on emotions.

By supporting a value-updating perspective, the current work also has much broader implications for affective science. Building on a core idea from this perspective—that emotions arise when there is a perceived need to update one's stored value representations—we are currently developing these implications into what we call the Value-Updating Theory of Emotion. For example, one implication of this view is that emotions may be intrinsically goal-directed processes aimed at resolving RPEs or updating values; moreover, common responses to emotions, such as crying or seeking social support when sad, may be understood as efforts to facilitate value updating, and thus help "process" or "resolve" the emotion. These ideas

offer a novel cognitive-computational perspective on what it means to "resolve" an emotion, and highlight deep connections between the mechanisms, function, and regulation of emotions. This perspective also highlights new directions for research on dysregulation and disorders of emotions. For example, it can provide an integrative view of why emotions sometimes become "stuck" in an unresolved state, as in PTSD or prolonged grief disorder: if conditions make value-updating seem necessary, yet make it difficult to achieve (e.g., due to being uncertain about how to update one's values, or unwilling to accept the implications of this update). Moreover, links to computational reinforcement learning may offer powerful tools for formalizing these processes and how they go awry—e.g., using model-based reinforcement learning (Doll et al., 2012) to account for how uncertainty tends to prolong emotions and make them difficult to resolve (Baum et al., 1997; Wilson & Gilbert, 2008).

Thus, to sum up, the current work supports one of the central predictions of a value-updating perspective on emotions regarding the key role of RPEs in emotion generation, and highlights how seemingly conflicting results can be reconciled with this view. In doing so, it aims to contribute to a long and fruitful line of research into the fundamental computations underlying emotions, and their broad implications for affective science.

Statements and Declarations

Competing Interests

The authors have no competing interests to disclose.

Compliance with Ethical Standards

The study was approved by the ethics committee of KU Leuven, and all participants gave informed consent before participating.

Data Availability

The preregistration, data, analysis code, and study materials are available at https://osf.io/v5ubg/?view_only=bf3e05ef57f74e988a9941892caadf18.

Appendix

Derivation of Mathematical Predictions

Building off the mathematical derivations in Parr et al (2026), here we show how we derived estimates for the additional contribution of outcomes and/or expectations per se, beyond RPEs, as well as how we computed bounds on the size of the trial-level RPE effect. For simplicity, we focus on cases where only trial-level outcomes, expectations and RPEs are relevant (e.g., because effects of block-level RPEs have faded away, and any other drivers of affect have been statistically controlled for). Suppose, in such cases, that the true model of affect is a linear combination of the effect of RPEs (i.e., outcome - expectation), as well as some potential additional effect of outcomes and/or expectations per se:

$$\text{Affect} = c_{RPE}\ \text{RPE} + c_{outcome}\ \text{outcome} + c_{expectation}\ \text{expectation}$$

Note that the coefficients that capture the true contribution of each factor are denoted with *c*, to distinguish these from the coefficients estimated in the regression models (denoted as *b* here and in the main text). Ideally, one would be able to fit the regression model that corresponds to this equation to get estimates for each of these three coefficients; however, this is not possible for the current task, since the inherent relationships between RPEs, outcomes and expectations mean that this model would have perfect collinearity and no unique solution. (This will apply to all tasks, unless they include another measure that captures the unique contribution of one or two of these variables, for instance, an independent measure of how expectations impact affect, from before outcomes have been learned, as in one study in Rutledge et al., 2014.) Thus, in the current work, rather than fitting a regression with all three variables, at most two out of three variables

could be included. The fitted coefficients then reflect some combination of the true underlying effects, with the exact nature of this combination depending on the variables included. For example, if one chooses to include only outcomes and expectations in the model, as in the current work, these coefficients can be derived as follows. Rewriting the true model in terms of outcomes and expectations, to match the regression model of interest, gives:

$$\text{Affect} = c_{RPE}\,(\text{outcome} - \text{expectation}) + c_{outcome}\,\text{outcome} + c_{expectation}\,\text{expectation}$$

$$\text{Affect} = (c_{RPE} + c_{outcome})\,\text{outcome} + (c_{expectation} - c_{RPE})\,\text{expectation}$$

Thus, the fitted regression coefficients for outcomes and expectations correspond to the following combinations of the underlying effects:

$$b_{outcome} = c_{RPE} + c_{outcome} \quad \text{(Equation 1)}$$

$$b_{expectation} = c_{expectation} - c_{RPE} \quad \text{(Equation 2)}$$

Thus, it is not possible to directly recover any of the true underlying effects (any of the *c*'s) individually.

Nevertheless, rearranging and combining these equations *does* allow us to get an estimate for $c_{outcome} + c_{expectation}$, that is, the total additional effect of outcomes and/or expectations, beyond their impact through RPEs, in terms of the sum of the two regression coefficients:

$$c_{RPE} = b_{outcome} - c_{outcome}$$

$$c_{RPE} = c_{expectation} - b_{expectation}$$

$$b_{outcome} - c_{outcome} = c_{expectation} - b_{expectation}$$

$$c_{expectation} + c_{outcome} = b_{\text{outcome}} + b_{expectation}$$

This was used in the main analyses to estimate the size of any additional contribution of these variables.

In addition, while it is not possible to directly compute the true RPE effect (c_{RPE}), limits on its possible values can be computed if we assume that the true impact of outcomes and expectations per se is non-negative (in line with typical predictions of value-based perspectives, which assume positive effects):

$$c_{outcome} \geq 0$$

$$c_{expectation} \geq 0$$

Rearranging and combining this with the equation for $b_{outcome}$ (Equation 1) gives:

$$b_{outcome} - c_{RPE} = c_{outcome}$$

$$b_{outcome} - c_{RPE} \geq 0$$

$$b_{outcome} \geq c_{RPE}$$

Similarly, for $b_{expectation}$ (Equation 2) we get:

$$b_{expectation} + c_{RPE} = c_{expectation}$$

$$b_{expectation} + c_{RPE} \geq 0$$

$$c_{RPE} \geq -b_{expectation}$$

In other words, $b_{outcome} \geq c_{RPE} \geq -b_{expectation}$. This provides bounds on the possible RPE effects that would be consistent with the regression results.

This assumption of non-negative values is also important to interpreting our main analysis that provided evidence for trial-level RPE effects. In particular, in that analysis, we took the positive effect of trial coherence (a proxy for trial-level outcomes) as evidence that thes outcomes had a positive effect on affect, and the negative effect of average block coherence (a proxy for trial-level expectations) as evidence that these expectations had a negative effect on affect. However, as demonstrated in Equations 1 and 2, these fitted coefficients ($b_{outcome}$ and $b_{expectation}$) are not pure reflections of the true underlying impact of these variables ($c_{outcome}$ and

$c_{expectation}$), since they reflect a combination of RPEs and outcomes or expectations per se. Nevertheless, by treating the true impact of outcomes and expectations pe se ($c_{outcome}$ and $c_{expectation}$) as non-negative, as is generally assumed by work in this area, the observed regression results can in fact be used as evidence for trial-level RPE effects. The clearest way to see this is by considering Equation 2: if $c_{expectation} \geq 0$, then the only way for $b_{expectation}$ to give a negative value, as observed in the current results, is through there being a positive impact of RPEs (c_{RPE}). In other words, trial-level RPEs are the only theoretically-relevant way to get this negative expectation effect, if the impacts of expectations per se are otherwise assumed to be positive, confirming the role of trial-level RPEs in driving the current results.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Bennett, D., Davidson, G., & Niv, Y. (2022). A model of mood as integrated advantage. *Psychological Review*, *129*(3), 513.

Bhatia, S., Mellers, B., & Walasek, L. (2019). Affective responses to uncertain real-world outcomes: Sentiment change on Twitter. *PLOS ONE*, *14*(2), e0212489. https://doi.org/10.1371/journal.pone.0212489

Blain, B., & Rutledge, R. B. (2020). Momentary subjective well-being depends on learning and not reward. *eLife*, *9*, e57977. https://doi.org/10.7554/eLife.57977

Brandstätter, E., & Kriz, W. C. (2001). Hedonic Intensity of Disappointment and Elation. *The Journal of Psychology*, *135*(4), 368–380. https://doi.org/10.1080/00223980109603705

Broekens, J. (2018). *A Temporal Difference Reinforcement Learning Theory of Emotion: Unifying emotion, cognition and adaptive behavior* (arXiv:1807.08941). arXiv. https://doi.org/10.48550/arXiv.1807.08941

Buttrick, N., Axt, J., Ebersole, C. R., & Huband, J. (2020). Re-assessing the incremental predictive validity of Implicit Association Tests. *Journal of Experimental Social Psychology*, *88*, 103941.

Carver, C. S. (2015). Control Processes, Priority Management, and Affective Dynamics. *Emotion Review*, *7*(4), 301–307. https://doi.org/10.1177/1754073915590616

Chater, N., & Loewenstein, G. (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization, Thriving through Balance*, *126*, 137–154.

https://doi.org/10.1016/j.jebo.2015.10.016

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

Ding, K., Lin, H., Liu, G., Kong, F., Liu, J., & Zhou, X. (2025). The expectation-updating mechanism in gratitude: A predictive coding perspective. *Emotion*, *25*(1), 198–209. https://doi.org/10.1037/emo0001421

Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications*, *6*(1), 1–10.

Eldar, E., Roth, C., Dayan, P., & Dolan, R. J. (2018). Decodability of reward learning signals predicts mood fluctuations. *Current Biology*, *28*(9), 1433-1439. e7.

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences*, *20*(1), 15–24.

Emanuel, A., & Eldar, E. (2023). Emotions as computations. *Neuroscience & Biobehavioral Reviews*, *144*, 104977. https://doi.org/10.1016/j.neubiorev.2022.104977

Forbes, L., & Bennett, D. (2024). The effect of reward prediction errors on subjective affect depends on outcome valence and decision context. *Emotion*, *24*(3), 894–911. https://doi.org/10.1037/emo0001310

Gopnik, A. (2000). *Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory formation system.*

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife*, *5*, e13388. https://doi.org/10.7554/eLife.13388

Harrell Jr, F. E. (2025). *rms: Regression Modeling Strategies* (Version 7.0-0) [R package]. https://CRAN.R-project.org/package=rms

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363.

Keren, H., Zheng, C., Jangraw, D. C., Chang, K., Vitale, A., Rutledge, R. B., Pereira, F., Nielson, D. M., & Stringaris, A. (2021). The temporal representation of experience in subjective mood. *eLife*, *10*, e62051. https://doi.org/10.7554/eLife.62051

Krupić, D., & Corr, P. J. (2014). Individual differences in emotion elicitation in university examinations: A quasi-experimental study. *Personality and Individual Differences*, *71*, 176–180. https://doi.org/10.1016/j.paid.2014.08.001

Kurzban, R. (2016). The sense of effort. *Current Opinion in Psychology, Evolutionary Psychology*, *7*, 67–70. https://doi.org/10.1016/j.copsyc.2015.08.003

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

Lenth, R. (2019). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version 1.3.5) [R package]. https://CRAN.R-project.org/package=emmeans

Loomes, G., & Sugden, R. (1986). Disappointment and Dynamic Consistency in Choice under Uncertainty. *The Review of Economic Studies*, *53*(2), 271–282. https://doi.org/10.2307/2297651

Marshall, M., & Brown, J. (2006). Emotional reactions to achievement outcomes: Is it really best to expect the worst? *Cognition and Emotion*, *20*(1), 43–63. https://doi.org/10.1080/02699930500215116

McGraw, A. P., Mellers, B., & Tetlock, P. E. (2005). Expectations and emotions of Olympic athletes. *Journal of Experimental Social Psychology*, *41*(4), 438–446. https://doi.org/10.1016/j.jesp.2004.09.001

Mellers, B., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision Affect Theory: Emotional Reactions to the Outcomes of Risky Options. *Psychological Science*, *8*(6), 423–429. https://doi.org/10.1111/j.1467-9280.1997.tb00455.x

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

Mendl, M., Burman, O. H. P., & Paul, E. S. (2010). An integrative and functional framework for the study of animal emotion and mood. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1696), 2895–2904. https://doi.org/10.1098/rspb.2010.0303

Moors, A., Van de Cruys, S., & Pourtois, G. (2021). Comparison of the determinants for positive and negative affect proposed by appraisal theories, goal-directed theories, and predictive processing theories. *Current Opinion in Behavioral Sciences*, *39*, 147–152. https://doi.org/10.1016/j.cobeha.2021.03.015

Neville, V., Dayan, P., Gilchrist, I. D., Paul, E. S., & Mendl, M. (2021). Dissecting the links between reward and loss, decision-making, and self-reported affect using a computational approach. *PLOS Computational Biology*, *17*(1), e1008555. https://doi.org/10.1371/journal.pcbi.1008555

Otto, A. R., & Eichstaedt, J. C. (2018). Real-world unexpected outcomes predict city-level mood states and risk-taking behavior. *PLOS ONE*, *13*(11), e0206923. https://doi.org/10.1371/journal.pone.0206923

Parr, D., Bao, J., Madlon-Kay, S., Samanez-Larkin, G. R., & LaBar, K. S. (2026). *Affective*

valence tracks value rather than value updates during classical conditioning (Rekyz_v2). PsyArXiv. https://osf.io/preprints/psyarxiv/rekyz_v2/

R Core Team. (2024). *R: A language and environment for statistical computing.* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raz, I., Reggev, N., & Gilead, M. (2024). Is it better to be happy or right? Examining the relative role of the pragmatic and epistemic imperatives in momentary affective evaluations. *Emotion*, *24*(6), 1343–1357. https://doi.org/10.1037/emo0001349

Rutledge, R. B., Moutoussis, M., Smittenaar, P., Zeidman, P., Taylor, T., Hrynkiewicz, L., Lam, J., Skandali, N., Siegel, J. Z., Ousdal, O. T., Prabhu, G., Dayan, P., Fonagy, P., & Dolan, R. J. (2017). Association of Neural and Emotional Impacts of Reward Prediction Errors With Major Depression. *JAMA Psychiatry*, *74*(8), 790–797. https://doi.org/10.1001/jamapsychiatry.2017.1713

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252–12257.

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2015). Dopaminergic Modulation of Decision Making and Subjective Well-Being. *Journal of Neuroscience*, *35*(27), 9811–9822. https://doi.org/10.1523/JNEUROSCI.0702-15.2015

Schmidhuber, J. (2008). Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Workshop on Anticipatory Behavior in Adaptive Learning Systems*, 48–76.

Sharot, T., Rollwage, M., Sunstein, C. R., & Fleming, S. M. (2023). Why and When Beliefs

Change. *Perspectives on Psychological Science*, *18*(1), 142–151. https://doi.org/10.1177/17456916221082967

Shepperd, J. A., & McNulty, J. K. (2002). The Affective Consequences of Expected and Unexpected Outcomes. *Psychological Science*, *13*(1), 85–88. https://doi.org/10.1111/1467-9280.00416

Smith, C. A., & Lazarus, R. S. (1993). Appraisal components, core relational themes, and the emotions. *Cognition and Emotion*, *7*(3–4), 233–269. https://doi.org/10.1080/02699939308409189

Spector, A. J. (1956). Expectations, fulfillment, and morale. *The Journal of Abnormal and Social Psychology*, *52*(1), 51–56. https://doi.org/10.1037/h0047881

Topolinski, S., & Reber, R. (2010). Gaining Insight Into the "Aha" Experience. *Current Directions in Psychological Science*, *19*(6), 402–405. https://doi.org/10.1177/0963721410388803

Van de Cruys, S. (2017). Affective value in the predictive mind. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. MIND Group.

Vanhasbroeck, N., Devos, L., Pessers, S., Kuppens, P., Vanpaemel, W., Moors, A., & Tuerlinckx, F. (2021). Testing a computational model of subjective well-being: A preregistered replication of Rutledge et al. (2014). *Cognition and Emotion*, *35*(4), 822–835. https://doi.org/10.1080/02699931.2021.1891863

Verinis, J. S., Brandsma, J. M., & Cofer, C. N. (1968). Discrepancy from expectation in relation to affect and motivation: Tests of McClelland's hypothesis. *Journal of Personality and Social Psychology*, *9*(1), 47–58. https://doi.org/10.1037/h0025672

Villano, W. J., Otto, A. R., Ezie, C. E. C., Gillis, R., & Heller, A. S. (2020). Temporal dynamics

of real-world emotion are more strongly linked to prediction error than outcome. *Journal of Experimental Psychology: General*, *149*(9), 1755–1766. https://doi.org/10.1037/xge0000740

Vinckier, F., Rigoux, L., Oudiette, D., & Pessiglione, M. (2018). Neuro-computational account of how mood fluctuations arise and affect decision making. *Nature Communications*, *9*(1), 1708. https://doi.org/10.1038/s41467-018-03774-z

Voodla, A., Uusberg, A., & Desender, K. (2024). Affective valence does not reflect progress prediction errors in perceptual decisions. *Cognitive, Affective, & Behavioral Neuroscience*, *24*(1), 60–71. https://doi.org/10.3758/s13415-023-01147-8

Voodla, A., Uusberg, A., & Desender, K. (2025). Metacognitive confidence and affect – two sides of the same coin? *Cognition and Emotion*, *39*(8), 1857–1874. https://doi.org/10.1080/02699931.2025.2451795

Zeelenberg, M., van Dijk, W. W., Manstead, A. S. R., & vanr de Pligt, J. (2000). On bad decisions and disconfirmed expectancies: The psychology of regret and disappointment. *Cognition and Emotion*, *14*(4), 521–541. https://doi.org/10.1080/026999300402781

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*. https://doi.org/10.3389/fnint.2012.00079