Revisiting the electrophysiological correlates of valence and expectancy in reward processing – A Multi-lab replication

Katharina Paul[1†], Douglas J. Angus[2], Florian Bublatzky[3], Raoul Dietrich[4], Tanja Endrass[4], Lisa-Marie Greenwood[5], Greg Hajcak[6], Bradley N. Jack[5], Sebastian P. Korinth[7,8], Leon O. H. Kroczek[9], Boris Lucero[10], Annakarina Mundorf[11], Sophie Nolden[12], Jutta Peterburs[11], Daniela M. Pfabigan[13,14], Antonio Schettino[15,16], Mario C. Severo[17], Yee Lee Shing[8,12], Gözem Turan[8,12], Melle J. W. van der Molen[17], Matthias J. Wieser[18], Niclas Willscheid[3], Faisal Mushtaq[19], Yuri G. Pavlov[20], Gilles Pourtois[21]

[1]Faculty of Psychology and Human Movement Science, University of Hamburg, Hamburg, Germany, [2]School of Psychology, Bond University, Gold Coast, Australia, [3]Central Institute of Mental Health Mannheim, Medical Faculty Mannheim/Heidelberg University, Mannheim, Germany, [4]Faculty of Psychology, Technical University Dresden, Dresden, Germany, [5]Research School of Psychology, Australian National University, Canberra, Australia, [6]Department of Psychology and Department of Biomedical Sciences, Florida State University, USA, [7]DIPF, Leibniz Institute for Research and Information in Education Frankfurt am Main, Frankfurt am Main, Germany, [8]Center for Individual Development and Adaptive Education of Children at Risk (IDeA) Frankfurt am Main, Germany, [9]Department of Psychology, Clinical Psychology and Psychotherapy, University of Regensburg, Regensburg, Germany, [10]The Neuropsychology and Cognitive Neurosciences Research Center (CINPSI Neurocog), Faculty of Health Sciences, Catholic University of the Maule (UCMaule), Talca, Chile., [11]Institute of Systems Medicine & Department of Human Medicine, MSH Medical School Hamburg, Hamburg, Germany, [12]Department of Psychology, Goethe University Frankfurt, Frankfurt, Germany, [13]Department of Behavioural Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway, [14]Department of Medicine, Vestfold Hospital Trust, Tønsberg, Norway, [15]Erasmus Research Services, Erasmus University Rotterdam, Rotterdam, The Netherlands, [16]Institute for Globally Distributed Open Research and Education (IGDORE), Sweden, [17]Institute of Psychology, Leiden University, Leiden, The Netherlands, [18]Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands, [19]School of Psychology, University of Leeds, Leeds, United Kingdom, [20]Institute of Medical Psychology and Behavioral Neurobiology, University of Tuebingen, Tuebingen, Germany, [21]Department of Experimental Clinical & Health Psychology, Ghent University, Ghent, Belgium

[†]Correspondence should be addressed to Katharina Paul; E-mail: katharina.paul@gmx.at

Author contributions:

Author contributions are coded according to the CRediT taxonomy (Allen et al., 2014)

Annakarina Mundorf: Investigation and Writing - review & editing.
Antonio Schettino: Data curation, Formal analysis, Methodology, Software, Validation, and Writing - review & editing.
Boris Lucero: Funding acquisition and Investigation.
Bradley N. Jack: Funding acquisition, Investigation, Supervision, and Writing - review & editing.
Daniela M. Pfabigan: Funding acquisition, Methodology, Software, Supervision, and Writing -

review & editing.
Douglas J. Angus: Supervision and Writing - review & editing.
Faisal Mushtaq: Conceptualization, Project administration, and Writing - review & editing.
Florian Bublatzky: Funding acquisition, Supervision, and Writing - review & editing.
Gilles Pourtois: Conceptualization, Funding acquisition, Project administration, Supervision, Writing - original draft, and Writing - review & editing.
Gözem Turan: Investigation and Writing - review & editing.
Greg Hajcak: Writing - review & editing.
Jutta Peterburs: Funding acquisition, Methodology, Supervision, and Writing - review & editing.
Katharina Paul: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing - original draft, and Writing - review & editing.
Leon O. H. Kroczek: Investigation, Supervision, and Writing - review & editing.
Lisa-Marie Greenwood: Investigation, Supervision, and Writing - review & editing.
Mario C. Severo: Formal analysis, Validation, Visualization, and Writing - review & editing.
Matthias J. Wieser: Investigation and Supervision.
Melle J. W. van der Molen: Supervision and Writing - review & editing.
Niclas Willscheid: Investigation and Writing - review & editing.
Raoul Dietrich: Investigation, Methodology, and Writing - review & editing.
Sebastian P. Korinth: Investigation.
Sophie Nolden: Supervision and Writing - review & editing.
Tanja Endrass: Supervision and Writing - review & editing.
Yee Lee Shing: Funding acquisition, Supervision, and Writing - review & editing.
Yuri G. Pavlov: Conceptualization, Funding acquisition, Project administration, and Writing - review & editing.

Abstract

Two event-related brain potential (ERP) components, the frontocentral feedback-related negativity (FRN) and the posterior P300, are key in feedback processing. The FRN typically exhibits greater amplitude in response to negative and unexpected outcomes, whereas the P300 is generally more pronounced for positive outcomes. In an influential ERP study, Hajcak et al., (2005) manipulated outcome valence and expectancy in a guessing task. They found the FRN was larger for negative outcomes regardless of expectancy and the P300 larger for unexpected outcomes regardless of valence. These findings challenged the dominant Reinforcement Learning Theory of the ERN. We aimed to replicate these results within the #EEGManyLabs project (Pavlov et al., 2021) across thirteen labs. Our replication, including robustness tests, a PCA and Bayesian models, found that both FRN and P300 were significantly modulated by outcome valence and expectancy: FRN amplitudes (no-reward - reward) were largest for unexpected outcomes, and P300 amplitudes were largest for reward outcomes. These results were consistent across different methods and analyses. Although our findings only partially replicate the original study, they underscore the complexity of feedback processing and demonstrate how aspects of Reinforcement Learning Theory may apply to the P300 component, reinforcing the need for rigorous ERP research methodologies.

Highlights

- Large-scale replication with 359 participants across 13 labs worldwide.
- FRN during feedback processing is modulated by both expectancy and valence.
- P300 during feedback processing is modulated by both expectancy and valence.
- Stringent methods: preregistration, robustness tests, meta-analysis, PCA.

1. Introduction

Performance monitoring is critical for detecting possible mismatches between goals and actions and, upon their detection, triggering specific remedial processes (Ullsperger, Fischer, et al., 2014). This monitoring can be based either on internal cues, such as response errors, or external ones, such as unfavorable or negative evaluative feedback. A wealth of studies has used electroencephalographic (EEG) methods in humans and established the electrophysiological correlates of performance monitoring when it is based on internal or external cues (Ullsperger, Danielmeier, et al., 2014). Regarding the latter process, two distinct and successive event-related potential (ERP) components have been identified as reliable markers of performance monitoring: the feedback-related negativity (FRN) (Gehring & Willoughby, 2002) and the P300 (Courchesne et al., 1977). The FRN is a negative component recorded at fronto-central electrodes along the midline (most pronounced at electrodes Fz and FCz) that typically peaks around 250 ms after feedback onset. It is larger (i.e., more negative-going) for negative than positive feedback/outcomes (Miltner et al., 1997). Following the FRN, the P300 component, or more specifically the P3b (Polich, 2007; Walentowska et al., 2016), is elicited around 300-500 ms following feedback onset, and shows a more central/centro-parietal scalp distribution than the FRN (electrodes Cz and Pz). The P300 is larger (i.e., more positive-going) for unexpected/infrequent than expected/frequent events (Johnson & Donchin, 1980; Polich, 2007). The P300 is most often studied in the context of attention (Herrmann & Knight, 2001) and might reflect motivational processes involved during outcome and feedback processing (Huvermann et al., 2021; San Martín, 2012). Along these lines, these two ERP components likely reflect different aspects of information processing and/or a progressive accumulation of evidence of internal predictions endorsed by the participant during performance monitoring (Ullsperger, Danielmeier, et al., 2014).

The influential Reinforcement Learning Theory of the ERN (ERN-RL) put forward by Holroyd and Coles (2002) proposed that the FRN (and its response-based counterpart, the error-related negativity (ERN, Gehring et al., 2018) is a scalp manifestation of neural activity originating from the (dorsal) ACC, which itself receives direct dopaminergic inputs from the basal ganglia, including the striatum. In this model (Holroyd & Coles, 2002; see also Nieuwenhuis et al., 2004), the FRN reflects the detection of a discrepancy between the actual and the expected outcome (i.e., prediction error). Moreover, the FRN appears to be somewhat monotonically related to the size of the prediction error: the more unexpected an outcome is, the larger is the FRN (Holroyd et al., 2009; Weismüller & Bellebaum, 2016), although this relationship might not be linear (Williams et al., 2017). Whether the feedback is utilitarian (e.g., incentive-related) or performance-related (e.g., informing about accuracy) is irrelevant, as this prediction error captured by the FRN is equally large for unexpected outcomes in both cases (Nieuwenhuis, 2004).

Using this framework, Hajcak et al. (2005) performed an EEG study in which they assessed amplitude changes of the FRN and P300 components as a function of both valence and expectancy. They used a guessing task (a.k.a. the Doors Task; see Holroyd et al., 2003) in which participants had to guess which of four presented doors hid a small monetary prize (0.10$ reward). Importantly, prior to the choice, the probability to win (25%, 50%, or 75%) was announced to manipulate outcome expectancy. Results showed that the FRN did not differentiate between

these three levels of expectancy, while the P300 increased as a function of unexpectedness (i.e., it was more pronounced for unexpected (25%) than neutral (50%) outcomes, and for neutral than expected (75%) outcomes). These findings were found across two experiments in which expectancy was manipulated trial-wise (N = 17) and block-wise (N = 12), respectively.

In the following years, these findings received mixed support, and the extent to which the P300 is insensitive to valence and the FRN is insensitive to expectancy remains contested. Whereas various experiments and meta-analyses have consistently shown that the P300 increases with outcome unexpectedness (Stewardson & Sambrook, 2020), the effect of outcome valence on the P300 remains unclear. Some studies report similar results as Hajcak et al. (2005), i.e., no effect of outcome valence on the P300 component (Pfabigan et al., 2011), yet others have shown effects in the opposite direction, i.e., positive outcomes elicited either larger or smaller P300 amplitudes (Glazer et al., 2018; San Martín, 2012; Stewardson & Sambrook, 2020). To explain these discrepancies, methodological differences such as imbalanced stimulus frequencies, have sometimes been discussed (Stewardson & Sambrook, 2020). In comparison, the observed insensitivity of the FRN to expectancy has gained much more attention as this observation was at odds with the predictions of the ERN-RL theory (Holroyd & Coles, 2002; Walsh & Anderson, 2012) and inconsistent with previous empirical observations (Holroyd et al., 2003).

To reconcile the divergent findings, Hajcak et al. (2005) suggested that this signed prediction error effect conferred to the FRN was observed using trial-and-error learning tasks, as opposed to guessing tasks. Consistent with this interpretation, later ERP studies using learning-based tasks reported modulations of the FRN by expectancy (e.g., Ferdinand et al., 2012; Gu et al., 2021; Holroyd et al., 2009; Warren & Holroyd, 2012), while expectancy modulations were only rarely found in guessing tasks (Gheza et al., 2018; HajiHosseini et al., 2012). The close coupling of choices, expectations, and the following outcomes could be at the core of this discrepancy (Hajcak et al., 2007). Thus, while this finding for the FRN was surprising at first, subsequent studies and some meta-analyses confirmed that insensitivity (or lower sensitivity) of the FRN to expectancy could be common in contexts in which learning remains inherently limited, such as in guessing tasks (e.g., Guthrie, 1942; Sambrook et al., 2012).

This original study has engendered a large amount of ERP studies and theoretical models, which have often used similar guessing tasks, and characterized the electrophysiological correlates of reward processing during performance monitoring in various contexts and situations (see Glazer et al., 2018; San Martín, 2012; Walsh & Anderson, 2012). Moreover, following the publication of this study, several methodological and theoretical refinements have been proposed to explore reward-based feedback processing at the FRN level. Chief amongst these developments has been the recognition that variation in the FRN signal may be the product of a superimposed positive-going deflection, a so-called Reward Positivity (RewP; see Proudfit, 2015). When conceptualizing feedback-related ERPs as the difference between positive and negative outcomes, the component labels are interchangeable as this new perspective affects only the direction of the effects (i.e., for unexpected outcomes the component is more positive or more negative) (Krigolson, 2018; Proudfit, 2015). However, when looking at the condition-specific ERPs, this new perspective affects the sign of the prediction error. If the response to negative, “worse-than-expected”, outcomes drives the effects, the FRN/RewP captures a negative prediction error. If the response to positive, “better-than-expected”, outcomes drives the effects,

the FRN/RewP captures a positive prediction error. While many attempts have been made to disentangle these different responses (Foti et al., 2011; Gable et al., 2021; Gheza et al., 2018), the FRN/RewP probably captures both due to the underlying frequency responses (Bernat et al., 2015; Hoy et al., 2021). Nevertheless, this paradigm shift did not only move the focus towards positive (as opposed to negative) outcomes, but also contributed to important methodological discussions about how to best measure this early ERP component following feedback onset (Klawohn et al., 2020). Hence, it appears important to investigate if the sensitivity to expectedness is driven by the response to positive or negative outcomes.

The results of this study sparked numerous conceptual replications on the nature of the FRN/RewP and the P300 component across different tasks, motivational contexts, and in clinical and non-clinical populations. To date, the work has been cited over 620 times (Google Scholar in November 2024). Yet, despite this intense focus, there has been no direct replication of the original procedure, measures, and analyses. The goal of the present study was to undertake a multi-lab replication of Hajcak et al. (2005), using a trial-by-trial manipulation of both expectancy and valence. We intended to complement this direct replication with modern preprocessing and analytical approaches to test the robustness of the reported effects. Based on Hajcak et al. (2005), we hypothesized that:

1. The FRN/RewP will not vary with expectancy. More specifically, the amplitude of the FRN/RewP will not be statistically different for expected, neutral, and unexpected outcomes.
2. The amplitude of the P300 will increase as a function of unexpectedness (i.e., unexpected > neutral > expected), irrespective of valence (reward vs. no-reward).

Finally, if, in contrast to the original replication, but in line with the RL-Theory, we would find an effect of expectedness on FRN/RewP amplitudes, we would explore if this effect is driven by the response to reward or no-reward outcomes.

2. Methods

2.1. Statistical power and recruitment procedures

To guide a decision on sample size, the non-significant interaction of expectancy and location for the FRN/RewP component reported in Hajcak et al. (2005) was used. Not only is this the smallest reported effect, it is also the key theoretically relevant result. Unfortunately, the original paper did not report a complete set of statistical results ("$F(2,32) < 1$"), so estimates of the effect size of $\eta_p^2 = 0.059$[1] were only a rough overestimation of the true effect size. Additionally, there was no meta-analytical evidence readily available for this effect to compare this estimate. While a meta-analysis by Sambrook & Goslin (2015) reported an effect size of $d = 0.71$ for expectancy modulation of the FRN/RewP (equal to calculated $\eta_p^2 = 0.11$), it is important to note that this was aggregated across mostly learning tasks, and it is reasonable (and also discussed by Sambrook & Goslin (2015)) to assume that the effect size could be smaller in guessing tasks. While this could be considered an

[1] For this and the following statistics, η_p^2 was calculated from the reported F values (Cohen, 1988; Lakens, 2013), when no F values were reported, we used $F = 1$.

upper bound of the FRN/RewP effect of expectancy during guessing tasks, we refrained from using this estimate to guide an *a-priori* sample size determination.

To circumvent these limitations, we opted for a sensitivity analysis. Based on available resources, each of the thirteen replicating labs will provide the data from 25 participants (excluding participants because of computer malfunction, drop out, technical problems, or insufficient clean data (see below)), resulting in a sample size of 325 participants across all labs. With such a sample size, a sensitivity analysis in MorePower (6.0.4. Campbell & Thompson, 2012) showed that the smallest effect size that can be reliably detected is $\eta_p^2 = 0.014$ ($\alpha = .02$, $1 - \beta = .90$, 3x3 interaction in repeated measures ANOVA). This allowed us to identify a much smaller effect than any individual study on this matter has been able to identify so far.

A similar rationale was applied to the non-significant valence effect on the P300 ($F(1,16) = < 1$, calculated $\eta_p^2 = 0.048$) and the non-significant interaction of valence and expectancy ($F(2,32) = 2.88$, $p >. 09$, calculated $\eta_p^2 = 0.152$). In comparison, the effect size of the expectancy modulation on the P300 was reported to be relatively large ($F(2,32) = 45.48$, $p < .001$, $\mathcal{E} = .82$, calculated $\eta_p^2 = 0.740$). Even after dividing this effect size in half to correct for shrinkage effects commonly observed in replication studies (see Pavlov et al. (2021)), each individual lab had the statistical power to replicate this effect in the collected subsample ($n = 25$, $\alpha = .02$, $\eta_p^2 = 0.370$, $1 - \beta = .99$, main effect with 3 levels in repeated measures ANOVA).

In each replicating lab, participants were recruited via local advertisements or online recruitment systems. For their participation, they were reimbursed with 15 EUR/ NOK 200 or course credits. Additionally, each participant received a payout of their in-task wins of 5 EUR/ 17 AUD/ 50 NOK / 5000 CLP. Participants were told that they could increase their payouts if they chose the “correct door”. However, regardless of their choices the outcome was pre-programmed and unrelated to the choices made by the participants.

For each replicating lab (n=13), the study was approved by the local or national ethical committee/Institutional Review Board (ANU [2022/859]; Bond University [DA03365]; German Psychological Society (DGPS) [PK-22-02-21]; Ghent University [2022/14]; Leiden University [2022-05-12-M.J.W. van der Molen-V2-3819]; University of Bergen, Faculty of Psychology [2020/1926-28] & NSD [320122]; UCM [CEC-UCM 54/2023]; Erasmus University Rotterdam [ETH2223-0061]).

2.2. Procedure

The procedure followed the process employed in Experiment 1 in Hajcak et al. (2005) as closely as possible, and any departures from this were explicitly stated. Participants were tested individually in an EEG laboratory. Upon their arrival in the lab, they received a brief description of the experiment and provided informed consent. Then they were prepared for EEG recording and the EEG electrodes were attached. Participants were familiarized with the guessing task and the feedback using a practice block consisting of 40 trials (not included in the analysis). Afterwards, they completed 6 blocks of the guessing task, with each block comprising 40 trials (240 trials in total). Self-paced breaks were allowed in between blocks. Every other block, the experimenter

entered the testing room to inform about the current winnings (which were presented on the screen), monitored the EEG signal, and kept participants alert.

As this project was part of a wider initiative on replicability in EEG (#EEGManyLabs), most of the laboratories in this replication also collected resting state data EEG data together with some personality measures (https://osf.io/sp3ck/, (Pavlov et al., 2021). Neither EEG nor personality data was analyzed in the current study but will be merged across sites as part of a future replication project to be reported elsewhere. For this purpose, participating labs recorded 8 minutes of resting state EEG and participants will be asked to fill in three brief questionnaires (using previously validated translations into the local language where possible) prior to the start of the guessing task for the present study. These include the Karolinska Sleepiness Scale (KSS; Åkerstedt & Gillberg, 1990), the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) and the State Trait Anxiety Inventory Trait Version (STAI-T; Spielberger et al., 1970). After the guessing task, the labs recording this additional data asked participants to fill in the Edinburgh Handedness Inventory (EHI; Oldfield, 1971), the Behavioral Inhibition and Approach System Scales (BIS-BAS; Carver & White, 1994), the Center for Epidemiologic Studies Depression Scale (Radloff, 1977), and the Short Version of the Big Five Inventory (Gerlitz & Schupp, 2005) questionnaires. In the labs that did not record this additional data (see Supplementary Table 7), only the guessing task was presented.[2]

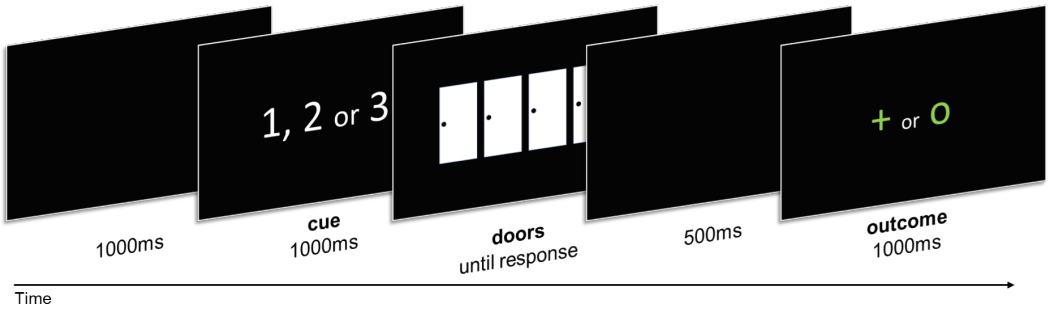


Figure 1: Trial structure. Each trial comprises three successive visual events: a cue (that informs about reward probability in the current trial), followed by the presentation of four doors (imperative stimulus; the participant is asked to pick one of them based on guessing), before the outcome (either reward or no-reward) is presented.

Each trial started with a cue presented for 1000 ms in the center of the screen (see Figure 1). The cue was presented as the number 1, 2, or 3, corresponding to a probability of winning of 25%, 50%, or 75% (i.e., how many of the four doors contained a prize). After this cue, four doors appeared in the center of the screen and the participant was asked to select one of them by pressing one of four predefined keys on the keyboard (exact keys varied across labs but correspond to four horizontally aligned keys pressed with the index and middle fingers of both hands, e.g., ZCBM for QWERTY keyboards, see Supplementary Table 7). Participants were

[2] Since the recording of the additional data before the guessing task took less than 15 minutes, we did not expect that these differences would affect the results. Nevertheless, we accounted for inter-lab variance in our statistical analyses (see below).

asked to guess which door could contain a prize. The four doors stayed on screen until the response/choice. Then a blank screen ensued (500 ms), before the outcome was presented in green font for 1000 ms. The outcome was presented as a "+", indicating that a small monetary reward was attained (value is 0.04 EUR or 0.15 AUD or 0.4 NOK or 35 CLP), or as a "o", indicating that no-reward was attained. The trial ended with a 1000 ms blank screen used as inter-trial interval. Stimuli were presented in white on black background. Accordingly, in this task, reward motivation was promoted while no punishment motivation was involved.

There were six experimental conditions, corresponding to the combinations of cue and outcome: expected reward (i.e., "+" symbol following "3" used as cue, 60 trials), neutral reward (i.e., "+" symbol following "2" used as cue, 40 trials), unexpected reward ("+" symbol following "1" used as cue, 20 trials), expected no-reward (i.e., "o" symbol following "1" used as cue, 60 trials), neutral no-reward (i.e., "o" symbol following "2" used as cue, 40 trials), and unexpected no-reward (i.e., "o" symbol following "3" used as cue, 20 trials). Across all blocks, these 6 conditions were shown in random order.

Upon completion of the task, participants were asked to answer two questions related to the attention paid to the numerical cue prior to the doors and the outcome during the experiment. These were answered on a seven-point scale, ranging from "ignored it" to "paid close attention" by the corresponding numbers on the keyboard.

The whole experiment lasted approximately 1 to 1.5 hours. The experiment was programmed using Presentation software (Neurobehavioral Systems, Inc., www.neurobs.com) and PsychoPy (Peirce, 2007) and translated into the local languages (English, Dutch, German, Norwegian, Spanish). Additional details on the used version of the experiment, the screen size, operating systems, used equipment etc. at each replicating lab are listed in the Supplementary Table 7.

2.3. Neurophysiological recordings

The replicating labs were using one of the following four EEG systems: (1) Biosemi Active 2; (2) BrainAmp DC, (3) BrainAmp actiCHamp Plus, (4) NeurOne Tesla. Using elastic caps, all labs recorded with either 32 or 64 channels positioned according to the extended 10/20 EEG system; (Chatrian et al., 1985)). One to four of these 32/64 electrodes or one to four additional external electrodes were used to record electro-oculogram (EOG), and two on the left and right mastoids. One EOG electrode was attached below the left eye, additional electrodes were placed above the left eye and on the outer canthi of the two eyes in some labs. The EEG (and EOG) data was sampled at 512, 500, 1000 Hz (depending on the setup). Labs also varied in their use of active vs. passive electrodes, and the applied online reference/ground (CMS/DRL, Cz, FCz, AFz). For details on each lab's set-up, see Supplementary Table 7.[3]

2.4. Artifact removal and EEG preprocessing

[3] The new recordings deviate from the original study in a few notable points: amplifier setup (Grass Model 7D polygraph with Neurosoft Quik-caps), number of recording sites (9), sampling rate (200 Hz), as well as pre-processing software (VPM) and applied offline filters (bandpass 0.05–35 Hz).

Data preprocessing closely followed the original study, including the following steps: activity recorded from Fz, Cz, and Pz and the additional external electrodes were: (i) re-referenced to Cz (the online-reference of the original study); (ii) filtered with a high-/low-pass filter of 0.05 and 35 Hz (the offline filter settings of the original study; EEGLAB defaults (Delorme & Makeig, 2004), transition band width 0.05/8.75 Hz, passband edge 0.05/35 Hz, cutoff frequency (-6dB) 0.025/39.38 Hz) (iii) down-sampled to 200/250/256 Hz as the original study recorded with a sampling rate of 200 Hz; (iv) segmented into epochs of interest (-500/+1500 ms around the onset of the outcome); (v) corrected for ocular artifacts (following Gratton et al., 1983, implemented into MATLAB); (vi) re-referenced to the linked mastoids; (vii) cleaned of segments containing artifacts (25 ms of invariant analog data on any channel; voltage exceeding ±100 μV)[4]; (viii) low-pass filtered at 20 Hz using a FIR filter (eeglab defaults, transition band width 5 Hz, passband edge 20 Hz, cutoff frequency (-6dB) 22.5 Hz); (ix) baseline corrected to -200 to 0 ms prior to outcome onset.

In addition to the use of a data preprocessing protocol that closely followed the one provided in the original study, the data was also preprocessed according to recent developments in psychophysiology, which allowed us to test the robustness of the results. Activity recorded from all EEG sensors was: (i) down-sampled to 500/512 Hz (if recorded with higher sampling rates); (ii) re-referenced to mastoids; (iii) high-pass filtered at 0.1 Hz using a FIR filter (eeglab defaults, transition band width 0.1 Hz, passband edge 0.1 Hz, cutoff frequency (-6dB) 0.05 Hz); (iv) low-pass filtered at 40 Hz using a FIR filter (eeglab defaults, transition band width 10 Hz, passband edge 40 Hz, cutoff frequency (-6dB) 45 Hz); (v) interpolated (spherically) if activity is invariant (>5 s) or not correlated to other channels (r<0.8); (vi) cleaned from bad segments identified by ASR (with burst criterion of 55 SD, ran on 1 Hz high-pass filtered data; segments flagged as bad are then removed from the unfiltered data); (vii) cleaned for ocular artifacts through an Independent Component Analysis (ICA, infomax, performed on 1Hz high-pass filtered data, rank lowered by the number of interpolated channels, otherwise eeglab defaults; weights were then applied to the unfiltered data) and ICLabel (based on the probability of being not a brain component (<30 %) but ocular artifacts (>70%)); (viii) segmented into epochs of interest (-200/+800 ms around the onset of the outcome); (ix) baseline corrected to -200 to 0 ms prior to outcome onset; and (x) cleaned of bad segments (epochs deviating more than 3.29 SD (Tabachnick & Fidell, 2007) from trimmed normalized means with respect to joint probability, kurtosis or the spectrum).

2.5. Outlier handling

The original study did not mention the use of any particular outlier criterion, and therefore for the direct replication the data from all participants was included.

Nevertheless, to test the robustness of the results, we aimed to ensure good data quality in two ways: First, from all complete recordings, we excluded participants who had more than 75% of trials rejected (i.e., only 60 trials out of the 240 trials used). Second, we excluded participants who had less than 8 trials per condition (as the FRN/RewP shows good internal

[4] The original study excluded data segments based on invariant data and/or A/D values exceeding the converter's minimum/maximum values. Since all replicating labs recorded with a different setup than the original study, we chose this cut-off instead.

consistency with at least 8 trials (Ethridge & Weinberg, 2018). Included trial number as well as standardized measurement error (Luck et al., 2021) were calculated and reported to describe data quality across conditions (and across participating labs).

To ensure that all participants paid attention to the numerical cues as well as the outcome, participants were excluded if they indicated in the attention ratings that they ignored the cue (i.e., answering with one or two on the seven-point scale).

2.6. Quantification of the ERPs

The FRN/RewP was quantified at Fz, Cz, and Pz as follows: First, a difference wave was created by subtracting the ERP observed for reward outcomes from the ERP observed for no-reward outcomes. This difference wave was computed separately for expected outcomes (expected no-reward minus expected reward), neutral outcomes (neutral no-reward minus neutral reward), and unexpected outcomes (unexpected no-reward minus unexpected reward). For each level of expectancy, the FRN/RewP was initially defined as the maximum negative amplitude of these difference waves within a window between 200 and 500 ms following outcome onset. This quantification procedure led to the peak of the FRN/RewP component to be misclassified with an average peak of 325 ms (*SD* = 87, *Range* = 203 - 496). In around 30% of cases, the FRN/RewP peak was identified *after* the P300 peak. We therefore repeated the analysis constraining the time window to end at the peak of the P300 component (if earlier than 500 ms after outcome onset). These results were mostly similar to the original quantification method. We report the results from the more appropriately scored FRN in the main text and highlight possible differences (where they arose) in the footnotes.

The P300 was scored at Pz as follows. Unlike the FRN/RewP, no difference wave was created. For each of the six conditions, the P300 was defined as the most positive peak in the ERP 200 to 600 ms following outcome onset.

In addition to this direct replication of the ERP components, we also scored the FRN/RewP and the P300 as mean amplitudes, since peak amplitude values are often more sensitive to high-frequency noise (Luck, 2014). Together with comparing different preprocessing of the data, this allowed us to test the robustness of the results. The FRN/RewP was scored following current recommendations as the mean amplitude 200-300 ms following outcome onset (Gheza et al., 2018; Krigolson, 2018; Proudfit, 2015; Sambrook & Goslin, 2015), while the P300 was scored as the mean amplitude 300-500 ms following outcome onset.

Moreover, since difference waves reduce some of the information helpful for follow-up tests, we additionally scored the FRN/RewP using the actual condition ERPs at Fz (for both peak and mean scoring).

Considering that the FRN/RewP and the P300 components occur in rapid succession, we additionally quantified the EEG components in terms of a principal component analysis (PCA) to ascertain possibly dissociable effects on these components and to disentangle them better using the ERP PCA Toolkit (EP Toolkit, version 2.80; Dien, 2010b). The individual ERPs (for each of the six conditions) from the preprocessing following current standards and after excluding outliers (see above) was used for this analysis. Considering the differences in the recording systems that were used, the individual ERPs were first standardized. Specifically, data was downsampled to a

common denominator (500 Hz) and only 56 electrodes which were common across most labs were used (8 labs, 224 participants)[5]. The ERPs were then subjected to a recommended two-step sequential PCA (Spencer et al., 1999, 2001). If not further specified, all default values in the graphical interface were used. The procedure began with a temporal Promax rotation to capture the variance across the time points from the average ERP data, followed by a spatial Infomax (ICA) rotation to obtain the variance of the spatial distribution of the data across the common recording sites (Dien, 2010a). The number of factors retained in each step depended on the scree plot, such that only factors explaining more variance than identified in random data was included (similar to parallel testing, see Dien, 2012). From all temporospatial factor combinations, default windowing was applied to screen out factors explaining less than 0.5% variance. All remaining factors were reconstructed into voltage space, in which the voltage accounted for at the peak time point and channel were evaluated as ERP waveforms. Factors whose peak latencies and channels coincided (based on visual inspection) with the canonical scalp distribution and time course of the FRN/RewP (fronto-central, 200-300 ms) and P3 components (posterior-central, 300-500 ms) were tested.

2.7. Statistical Analyses

The main focus of the analyses was (1) a direct replication of the approach applied in the original study using repeated measures analyses of variance (ANOVAs). However, we also tested the robustness of these effects (2) in multilevel models (MLMs), and (3) in a meta-analysis of our effects identified in each lab.

2.7.1. Direct Replication through ANOVAs

The ERP amplitudes calculated from the preprocessing and quantification methods used in the original study were subjected to two ANOVAs. For the FRN/RewP, the peak amplitude values were analyzed using a 3 (Location) x 3 (Expectancy) ANOVA. For the P300, a 2 (Valence) x 3 (Expectancy) ANOVA was used. In case a sphericity violation was detected, Greenhouse–Geisser correction was applied to *p* values. The significance alpha level was set to 0.02.

Moreover, to test if the results for the FRN/RewP were driven by the response to reward outcomes or no-reward outcomes, we calculated a 2 (Valence) x 3 (Expectancy) ANOVA on the amplitudes extracted at Fz (where it was shown to be maximal in the original study) together with the corresponding post-hoc tests. The main analyses are complemented by a series of robustness analyses (see below and Table 1).

2.7.2. Robustness test through MLMs

To better account for variability across participants and laboratories, we fitted eight Bayesian multilevel linear models on the FRN/RewP and P300 amplitude values. These models were set up identically, but the dependent variable was extracted either after (1) “original” or “current standard” preprocessing pipelines, and (2) quantified as either “peak” scores (as in the original publication) or as “mean” scores (as a more robust measure of the ERP components). By

[5] Restricting the analyses to only common channels across all thirteen labs resulted in a dramatically lower number of channels (19). Hence, we chose to include those channels present in most labs as a tradeoff between sample size and channel number.

crossing these analytical choices, we were able to assess the impact of these choices on the outcome and the robustness of the replication.

The models were specified as follows (in Wilkinson notation (Wilkinson & Rogers, 1973)):

FRN/RewP_amplitudes = 1 + location * expectancy + (1 + location * expectancy | laboratory / participant) [6]

P300_amplitudes = 1 + valence * expectancy + (1 + valence * expectancy | laboratory / participant)

Robustness test 1. Amplitudes were extracted after the preprocessing of the original publication and defined as the maximum peak in the specified time window. This followed the analysis of the original publication most closely, while controlling for inter-lab variance.

Robustness test 2. Amplitudes were extracted after the preprocessing of the original publication and defined as the mean in the specified time window.

Robustness test 3. Amplitudes were extracted after the preprocessing according to current standards and defined as the maximum peak in the specified time window.

Robustness test 4*.* Amplitudes were extracted after the preprocessing according to current standards and defined as the mean in the specified time window.

We allowed intercepts and slopes to vary as a function of participant and laboratory, to model varying effects on amplitude peak (or mean) originating from different laboratory setups and individual characteristics (e.g., skull thickness, hair). As a likelihood function, we chose a Gaussian distribution.

An important aspect of Bayesian analysis is the choice of priors (e.g., Natarajan & Kass, 2000). Given the unknown susceptibility of the electrophysiological signal to inter-individual differences in relation to the predictors of interest, we placed a weakly informative prior on intercepts and slopes: a normal distribution with $\mu = 0$ and and $\sigma = 10$. Since we had no prior knowledge regarding the other model parameters (e.g., standard deviation of laboratory or participant), we kept the software default weakly informative priors.

Models were fitted in *R* using the *brms* package (Bürkner, 2018), which employed the probabilistic programming language *Stan* (Carpenter et al., 2017) to implement a Markov chain Monte Carlo (MCMC) algorithm (Hoffman, 2014) to estimate posterior distributions of the parameters of interest. We started sampling by using 4 MCMC chains with 4000 iterations (2000 warm-up) and no thinning. In case of non-convergence, we increased the number of iterations by 500 until convergence was reached or a maximum of 8000 iterations per chain. Model convergence was assessed as follows: (i) visual inspection of trace plots, rank plots, and graphical posterior predictive checks (Gabry et al., 2019); (ii) Gelman-Rubin $\hat{R}$ statistic (Gelman & Shalizi, 2013)

[6] Additionally, we reported the following model in the supplement: FRN/RewP _amplitudes_at_Fz = 1 + valence * expectancy + (1 + valence * expectancy | laboratory / participant). This additional analysis helped to identify if the response to reward outcomes or no-reward outcomes was driving the effect.

between 1 and 1.05 (see also Nalborczyk et al., 2019). Goodness-of-fit was assessed via Bayesian R^2 (Gelman et al., 2019).

Posterior distributions of the model parameters were summarized using the mean and 95% credible interval (CI). Differences between conditions were calculated by computing the difference between posterior distributions of the respective conditions and summarized as above.

The existence of an effect was ascertained using the MAP-Based *p*-Value (*pMAP*), a Bayesian equivalent of the frequentist *p*-value (Mills, 2018). This index represents the odds of the posterior distribution of the parameter of interest against the point null hypothesis $H_0 = 0$ and, mathematically, corresponds to the density value at 0 divided by the density at the Maximum A Posteriori (MAP) (see also Makowski et al., 2019). Following the current arbitrary *p*-value convention for Registered Reports in Cortex, we considered an effect statistically significant if *pMAP* < .02.

Two caveats of the *pMAP* should be noted here (Makowski et al., 2019). First, just like the frequentist p-value, *pMAP* allows us to assess the *presence* of an effect, not its *magnitude* or *practical importance*. Second, *pMAP* is sensitive only to the amount of evidence for the *alternative hypothesis* H_1, but it is *not* useful when assessing the amount of evidence in favor of the *null hypothesis* H_0. In our case, *pMAP* < .02 would suggest that the effect is statistically significant. However, *pMAP* > .02 would not allow us to conclude that the effect does not exist, only uncertainty about its existence (absence of evidence rather than evidence of absence).

To address these issues and increase the informativeness of our results, we additionally computed Bayes factors (BFs; (Jeffreys, 1998; Kass & Raftery, 1995; Morey et al., 2016). BFs indicate *"the extent to which the data sway our relative belief from one hypothesis to the other"* (Etz & Vandekerckhove, 2018, p. 10). BFs were calculated as a Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers et al., 2010), i.e., comparing the marginal likelihoods of the alternative model against a model in which the tested parameter (i.e., the posterior distribution of condition differences) has been restricted to the point-null. We descriptively qualified BFs according to the arbitrary convention proposed by Kass & Raftery (1995): (i) $BF_{10} = 1$: *no* evidence in favor of H_1; (ii) $1 < BF_{10} < 3$: *weak* evidence in favor of H_1; (iii) $3 < BF_{10} < 20$: *positive* evidence in favor of H_1; (iv) $20 < BF_{10} < 150$: *strong* evidence in favor of H_1; (v) $BF_{10} > 150$: *very strong* evidence in favor of H_1. The reciprocal of BF_{10} (i.e., $BF_{01} = 1/BF_{10}$) indicated the corresponding evidence in favor of H_0.

As outlined, in our Bayesian multilevel models, we focused on estimating the posterior distributions of the parameters of interest rather than directly analyzing main effects and interactions. This approach provided a more nuanced understanding of the data by offering credible intervals for each parameter. As the model estimates the differences between specific conditions and their associated uncertainty, it is not designed to explicitly isolate and test interaction effects in the conventional sense.

2.7.3. Meta-Analysis (Robustness Test 5)

Even though each replicating lab only had the statistical power to test the effect of expectancy on the P300, the data of each lab was separately subjected to the same ANOVAs

described above (2.7.1). Then, a random effects meta-analysis was run where the effect sizes of valence (for the P300) or electrode (for the FRN/RewP), expectancy, and their interaction gathered in each replicating lab were combined. Following the method utilized previously in other large-scale replication projects (Ebersole et al., 2020; Open Science Collaboration, 2015), as implemented in the *esc* package for R (Lüdecke, 2019), we converted partial eta squared to correlation coefficients. Given that eta represents a non-directional effect size, we established directionality by fitting linear regression models, analogous to ANOVAs, and derived the sign of the regression coefficients for each effect of interest. We utilized Fisher's *z*-transformed correlation coefficients, adjusted by the signs from the linear regressions, from each laboratory in our meta-analyses. The back-transformed correlation coefficients are presented and depicted in forest and funnel plots. The *metafor* package (Viechtbauer, 2010) for R was used for the meta-analysis.

2.7.4. Temporospatial Principal Component Analysis (PCA) (Robustness Test 6)

The PCA factors were analyzed using the statistics function of the EP toolkit using all default parameters. The implemented ANOVAs are robust against violations of statistical assumptions. It included the following features: (i) trimmed means (cutting the outer quartiles) and winsorized covariances that protect against outliers; (ii) a bootstrapping routine (499,999 simulations, ran 11 times) that estimated the population distribution instead of assuming the normality of this distribution; and (iii) a Welch–James approximate degrees-of-freedom statistic that did not assume the homogeneity of error variance (Dien, 2010b). The robust 2x3 repeated-measures ANOVA included the within-subject factors Valence and Expectancy. The p-value was adjusted with the Bonferroni correction for multiple comparisons. Follow-up tests for significant interactions were reported. In case the interaction effect needed a better characterization of its source, the EP Toolkit implements a Dunn-Šidák post-hoc test.

The PCA identified 31 temporal factors x 5 spatial factors based on the Scree plot, generating a total of 155 temporospatial factor combinations. Using an automated windowing step, the factors were further sifted through a predetermined minimum 0.5% threshold for accounted variance. The remaining PCA factors after the windowing step were then visually inspected for further analysis. Factors that only resembled the FRN/RewP and P300 components, based on canonical time course and scalp topography, were subjected to the robust ANOVA test.

Similar to the results from the main analyses above, we expected for the factor corresponding to the FRN/RewP a significant main effect for valence (more factor negativity for no-reward outcomes), but no effect of expectancy or their interaction. In contrast, for the factor corresponding to the P300 component, we expected a significant effect of expectancy (more factor positivity for unexpected outcomes), but no effect of valence or their interaction.

2.8 Evaluation of the Replication and Robustness of Effects

The replication's success was mainly evaluated in the light of the outcomes of the ANOVAs (see 2.7.1) above): The FRN/RewP results were considered to be replicated successfully if the ANOVA showed a significant main effect of position (Fz > Pz), but no significant effect of expectancy or the interaction of expectancy and position. The P300 results were considered to be replicated

successfully if the ANOVA showed a significant main effect of expectancy (unexpected > expected), but no significant effect of valence or the interaction of expectancy and position.

However, going beyond the mere replication of the original study, we provided preliminary robustness tests by comparing these results to the outcomes of the MLMs (see (2.7.2.) above) and a PCA (see 2.7.4) above). If the MLMs and the PCA provided evidence for a similar pattern of results as (2.7.1), the effect was considered not only to be replicated but robust and, to some extent, independent of analytical choices. If the direct replication failed, i.e., significant effects were detected where none were expected, or expected effects did not reach significance, the MLMs were particularly important to conclude if the effects are present or not. If the pattern diverged across the robustness tests, possible sources of these discrepancies were discussed (with regard to preprocessing choices and/or quantification of the ERPs). Finally, the results of the MLM, (Robustness Test 1) were compared to the meta-analysis (see (2.7.3) above).

2.9. Analysis of ratings

The descriptive statistics for the subjective ratings pertaining to the attention paid to the cue and the feedback were reported (see Hajcak et al., 2005).

2.10. Sharing of Data and Code

Pre-processing steps were carried out using EEGLAB 2022.0 (Delorme & Makeig, 2004) implemented in MATLAB 2019, while statistical analyses were carried out in R (R-Core-Team, 2019). All experimental procedures, pre-processing scripts, analytical analyses are shared openly via the Open Science Framework (OSF, https://osf.io/2w9gy). All collected data will be made available online through GIN (https://gin.g-node.org/katpa/Paul_et_al_Cortex_ManyLabs_Hajcak05). This study is a registered report, the preregistered stage 1 manuscript can be accessed at https://osf.io/db4rs.

3. Results

The results of the direct replication as well as all robustness tests are summarized in Table 1.

Table 1. *Overview of Analyses and reported Results*

	Hajcak et al	Direct Replication	Robustness Test 1	Robustness Test 2	Robustness Test 3	Robustness Test 4	Robustness Test 5	Robustness Test 6
Pre-processing	Original	Original	Original	Original	Current standard	Current standard	Original	Current Standard
Outlier handling	None	None	None	None	Applied	Applied	None	Applied
Quantification of ERPs	Peak	Peak	Peak	Mean	Peak	Mean	Peak	PCA
Statistical Test	ANOVA	ANOVA	MLM	MLM	MLM	MLM	Meta-Analysis	ANOVA
N FRN	17	307	307	360	297	328	13/307	230
N P300	17	360	360	360	323	328	13/360	230
FRN Replication								
Expectancy	not sign. $\eta p^2 <$ 0.08**	sign. $\eta_p^2 = 0.08$ [0.05, 0.13]	++	FZ + PZ --	++	FZ + PZ --	sign. $r = 0.32$ [0.22, 0.42]	(1/1) sign. ♦
Location	sign. $\eta_p^2 = 0.34$*	sign. $\eta_p^2 = 0.34$ [0.28, 0.39]	++	++	++	++	sign. $r = 0.60$ [0.52, 0.66]	*n.r.*
Location x Expectancy	not sign. $\eta_p^2 < 0.02$**	sign. $\eta_p^2 = 0.02$ [0.01, 0.04]	*n.a.*	*n.a.*	*n.a.*	*n.a.*	sign. $r = 0.16$ [0.05, 0.27]	*n.r.*
P300 Replication								
Valence	not sign. $\eta_p^2 < 0.06$**	sign. $\eta_p^2 = 0.32$ [0.24, 0.39]	++	++	++	++	sign. $r = 0.59$ [0.49, 0.68]	(2/3) sign.
Expectancy	sign. $\eta_p^2 = 0.74$*	sign. $\eta_p^2 = 0.37$ [0.32, 0.42]	++	++	+++	++	sign. $r = 0.63$ [0.56, 0.69]	(2/3) sign.
Valence x Expectancy	not sign. $\eta_p^2 = 0.15$*	not sign. $\eta_p^2 = \leq 0.001$ [≤ 0.001, 0.02]	*n.a.*	*n.a.*	*n.a.*	*n.a.*	not sign. $r = 0.07$ [-0.04, 0.17]	not sign.

Note: For the original results, the direct replication and the meta-analysis (Robustness Test 5), the entries show the effect sizes along the applied analysis (η_p^2, r) together with their 95% CI. For the Bayesian statistics (Robustness test 1-4), the ++/– is a descriptive summary of the positive/negative evidence for H1 across the relevant paired comparisons to approximate the main effects. For the PCA (Robustness Test 6), the number of components capturing the respective ERP and showing that significant effect is reported.
N refers to sample size of the analysis. *As the original study did not report an effect size, these are deduced from the reported F-statistics and p-values. **For non-significant effects, no exact statistics were reported and these values reflect the largest effect size compatible with those.

n.r. refers to not relevant: PCA includes spatial components and need to be considered as such. *n.a*. refers to not applicable: Robustnesstests 1-4 were carried out using paired comparisons using Bayesian MLMs. ♦ Unlike in the original study, the PCA included the two factors outcome expectancy and valence, which showed a significant interaction.

3.1. Participants

In total, 370 participants were tested across the thirteen labs (*M* = 28.46, *SD* = 4.39, *Range* = 21 - 37). All participants gave written informed consent. Sixty-six percent were women. Across all labs, 4 recordings were incomplete (e.g., computer failing, fainting of participants, battery issues, recording issues) and 5 participants were excluded since data from the mastoids were too noisy. For the original pre-processing, 2 participants had less than one trial after data cleaning. For the preprocessing according to current standards, 14 participants had less than eight trials after data cleaning. FRN/RewP or P300 peaks could not be detected in at least one condition for 66 participants (for the additional analyses where reward and no-reward outcomes were analyzed separately, this number increased to 108). Nineteen participants were excluded from analysis since they reported to not have paid attention to the cue. The final number of participants can be found in Table 1.

3.2. Direct Replication through ANOVAs

The FRN/RewP component showed the expected frontocentral distribution peaking on average around 270 ms after feedback onset (*SD* = 41, *Range* = 203 - 495). The P300 component showed the expected central distribution peaking on average around 355 ms after feedback onset (*SD* = 72, *Range* = 203 - 598), see Figure 2.

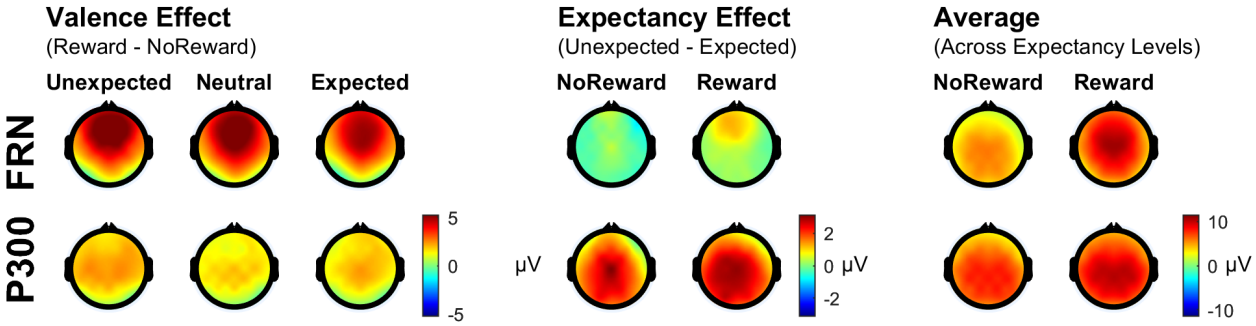


Figure 2. Topographical Plots of Valence and Expectancy Effects for FRN/RewP and P300 components.The FRN/RewP and P300 components were defined as the average amplitude in the 200-300 ms and 300-500 ms interval after outcome onset, respectively (preprocessing according to current standards as original preprocessing included only three channels).

Table 2. *Statistics of the direct replication*

	Effect	*F*	df_1	df_2	*p*	η_p^2	95 % CI η_p^2
FRN/RewP Component at Fz, Cz, Pz							
1	Expectancy	28.34	1.79	546.6	≤ .001	0.08	0.05, 0.13
2	Location	154.16	1.50	460.17	≤ .001	0.34	0.28, 0.39
3	Expectancy □ Location	6.71	2.65	811.27	≤ .001	0.02	0.01, 0.04
FRN/RewP Component at Fz							
4	Valence	514.64	1	262	≤ .001	0.66	0.60, 0.71
5	Expectancy	6.09	1.72	451.22	.004	.02	≤ 0.01, 0.05
6	Valence × Expectancy	10.70	1.81	474.82	≤ .001	0.04	0.01, 0.07
P300 Component at Pz							
7	Valence	167.73	1	359	≤ .001	0.32	0.24, 0.39
8	Expectancy	215.18	1.74	625.04	≤ .001	0.37	0.32, 0.42
9	Valence × Expectancy	1.60	1.72	617.4	.207	≤ 0.01	≤ 0.01, 0.02

The direct replication, using the original preprocessing and peak values, revealed for the FRN/RewP component significant main effects of Expectancy (table 2, row 1) and Location (table 2, row 2) and an interaction between these two factors (table 2, row 2). The FRN/RewP was largest for unexpected outcomes ($M_{Unexpected}$ = -9.65 µV, *sd* =6.46 vs. $M_{Neutral}$ = -9.03 µV, *sd* =5.6 vs. $M_{Expected}$ = -7.58 µV, *sd* =4.57) and at electrode Fz (M_{CZ} = -9.38 µV, *sd* =5.66 vs. M_{FZ} = -9.62 µV, *sd* =5.64 vs. M_{PZ} = -7.27 µV, *sd* =5.4). The difference between unexpected and expected outcome was largest at Fz (M_{CZ} = -1.79 µV, *sd* =6.27 vs. M_{FZ} = -2.75 µV, *sd* =6.28 vs. M_{PZ} = -1.81 µV, *sd* =5.91), see Figures 2, 3 and 4.[7]

To better understand the impact of expectancy and valence, we re-ran our analyses of the FRN/RewP component at Fz, treating reward and no-reward outcomes as separate conditions (i.e., without creating a difference wave prior to statistical analysis). This approach confirmed a significant main effect of Valence (Table 2, row 4), with more positive FRN/RewP amplitudes for reward compared to no-reward outcomes (M_{Reward} = 7.4 µV, *sd* =6.27 vs. $M_{NoReward}$ = 1.18 µV, *sd* =4.97). The main effect of Expectancy was also significant (table 2, row 5), with more positive values for unexpected compared to expected outcomes ($M_{Unexpected}$ = 4.67 µV, *sd* =7.22 vs. $M_{Neutral}$ = 3.88 µV, *sd* =6.2 vs. $M_{Expected}$ = 4.09 µV, *sd* =5.81). Additionally, the interaction between Valence and Expectancy was significant (Table 2, row 6), see Figure 4. The difference between unexpected and expected outcome was largest (most positive) and only significant for reward outcomes ($M_{NoReward}$ = -0.34 µV, *sd* =4.46, *p* = .18, *d* = -0.08 vs. M_{Reward} = 1.25 µV, *sd* =4.84, *p* < .001, *d* = 0.26). [8]

[7] When using the pre-registered quantification time window following the original study, similar results were obtained: Location ($F_{1.49,536.69}$ = 141.36, *p* ≤ .001, η_p^2 = 0.28, 95% CI [0.23, 0.33]), Expectancy ($F_{1.79,641.32}$ = 44.11, *p* ≤ .001, η_p^2 = 0.11, 95% CI [0.07, 0.15]), Interaction ($F_{2.8,1005.85}$ = 5.03, *p* = .002, η_p^2 = 0.01, 95% CI [≤ 0.001, 0.03])

[8] When using the pre-registered time window for quantifying the peak, the interaction was not significant: Valence ($F_{1,342}$ = 309.47, *p* ≤ .001, η_p^2 = 0.48, 95% CI [0.4, 0.54]), Expectancy ($F_{1.9,650.98}$ = 5.32, *p* = .006,

For the P300 component, the main effect of Valence was significant (Table 2, row 7), as was the main effect of Expectancy (Table 2, row 8). However, the interaction between these two factors was not significant (Table 2, row 9). P300 values were largest for reward compared to no-reward outcomes (M_{Reward} = 16 µV, *sd* =7.21 vs. $M_{NoReward}$ = 13.64 µV, *sd* =6.93), and for unexpected compared to expected outcomes ($M_{Unexpected}$ = 16.66 µV, *sd* =7.59 vs. $M_{Neutral}$ = 14.29 µV, *sd* =6.86 vs. $M_{Expected}$ = 13.51 µV, *sd* =6.65), see Figures 3 and 4.

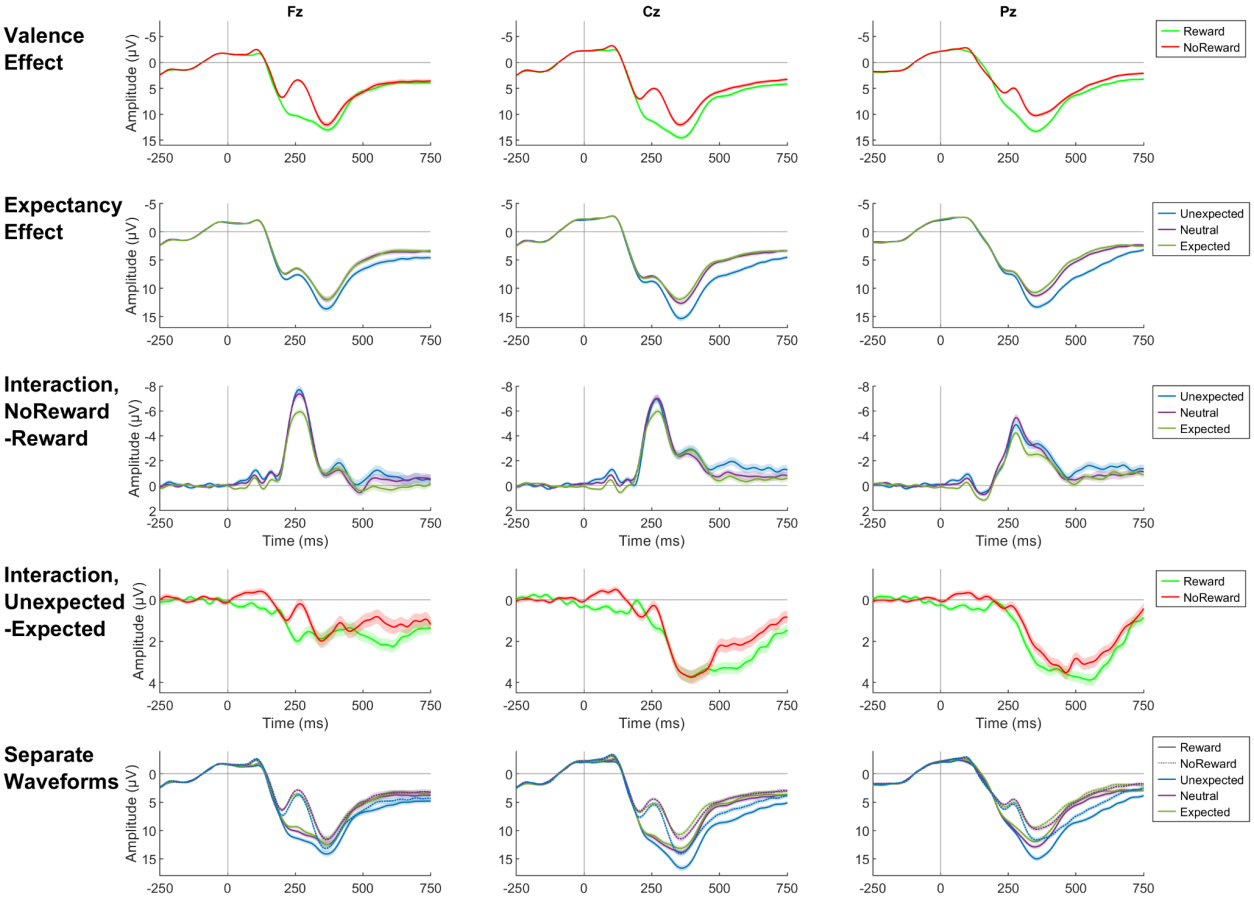


Figure 3. ERP Plots using the preprocessing following the original preprocessing at electrode sites Fz, Cz, and Pz, separately for the different conditions. Shaded Areas represent +- SEM.

η_p^2 = 0.02, 95% CI [≤ 0.001, 0.04]), Interaction ($F_{1.9,650.1}$ = 1.92, *p* = .149, η_p^2 = 0.01, 95% CI [≤ 0.001, 0.02]). Since determining a negative peak for reward outcomes can be difficult, using a mean window approach could provide a solution to score this component. However, using this alternative scoring method, the results were similar, see supplementary Table 3.

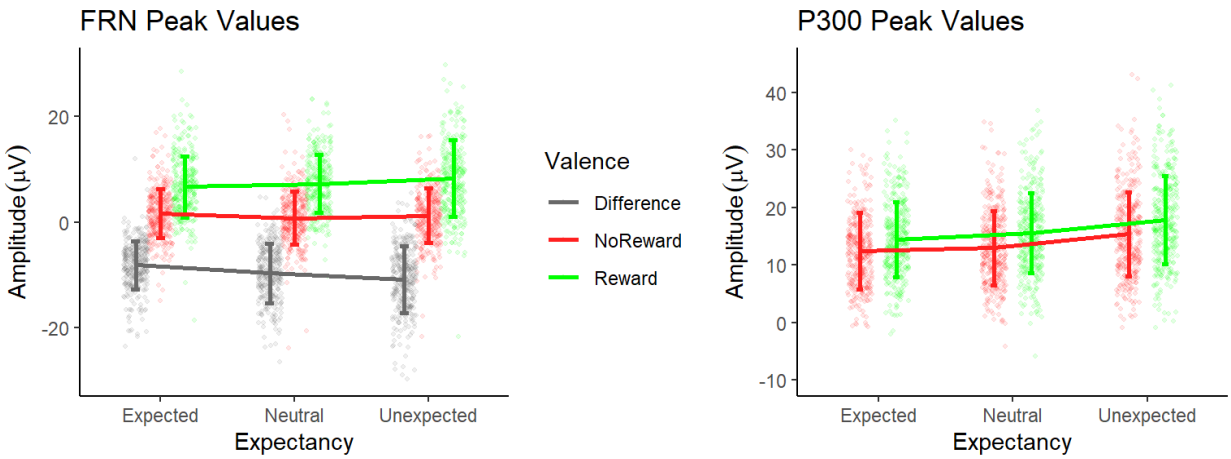


Figure 4. Mean ERP Values (+- SD) with individual data points for the direct replication, separately for each Expectancy and Valence level. The FRN component is shown at Fz, while the P300 component is shown at Pz.

3.3. Bayesian MLMs (Robustness test 1-4)

As we aimed to replicate a null effect, we included Bayesian statistics to allow testing for the absence of specific effects (i.e., Valence for P300, Expectancy for FRN). Moreover, to better control for unknown Lab effects in this multi-lab sample, we carried out these analyses using multilevel linear models. The results of the different Bayesian MLMs aligned with the results of the direct replication based on ANOVAs (see above). These robustness tests varied the preprocessing (original vs. current standard) as well as the quantification method (peak vs. mean amplitude). The detailed BF are reported in Tables 3 and 4. For more details, all comparisons, and parameters, please consult the supplementary section 8.3.

For the FRN/RewP component, we found positive evidence for the effect of Location across all preprocessing and quantification methods (Table 3, row 4/5, robustness test 1-4). In comparison, the interaction between Expectancy and Location was dependent on the quantification choice: When using a peak amplitude as quantification, there was positive to strong evidence for an effect of Expectancy at all electrodes (Table 3, row 1/2, robustness test 1/3). When using the mean amplitude as quantification (Table 3, row 1/2, robustness test 2/4), the Expectancy effect was only weakly supported at Fz (weak evidence for H_1), but not at Pz (positive evidence for H_0), suggesting that this effect could be robustly detected at electrode Fz, but not at Pz.

When we assessed the FRN/RewP component separately for reward and no-reward outcomes, we found strong evidence for the expected main effect of Valence, which was robustly found across all preprocessing and quantification methods (Table 3, row 7/8, robustness test 1-4). Regarding Expectancy, the type of quantification method used actually influenced the results: When using a mean quantification, we found positive evidence for an effect of Expectancy for reward outcomes (Table 3, row 5, robustness test 2/4), but not for no-reward outcomes (positive evidence for H_0, Table 3, row 6, robustness test 2/4). However, when using a peak quantification, the results were dependent on the preprocessing methods: For the original preprocessing, there was positive evidence for an effect of Expectancy for reward outcomes

(Table 3, row 5, robustness test 1), but not for no-reward outcomes (weak evidence for H_0, table 3, row 6, robustness test 1). In contrast, for the preprocessing according to current standards, the opposite pattern emerged: there was weak evidence against an effect of Expectancy for reward outcomes (Table 3, row 5, robustness test 3), but positive evidence for an Expectancy effect for no-reward outcomes (Table 3, row 6, robustness test 3). These results suggest that the mean quantification is probably better suited than the peak scoring to capture a robust effect of Expectancy for reward outcomes, given that they often do not elicit a clear peak (see last panel in Figure 3).

For the P300 component, we found positive to strong evidence for the main effect of Expectancy (BF_{10} = 14.84 - 28.76, p < .001) across all valence types, but also positive evidence for the main effect of Valence (BF_{10} = 4.72 - 8.85, p < .001) across all expectancy types. This pattern was robustly found across all preprocessing and quantification methods.

Table 3: *Bayes Factor Analysis for the different Robustness Tests and FRN/RewP component*

			Original		Current Standards	
			Peak (RobTest 1)	Mean (RobTest 2)	Peak (RobTest 3)	Mean (RobTest 4)
1	*Expectancy at Fz*	Unexpected Diff Fz - Expected Diff Fz	<.001* BF=12.5 ++	.004* BF=1.91 +	<.001* BF=23.15 +++	.004* BF=1.42 +
2	*Expectancy at Pz*	Unexpected Diff Pz - Expected Diff Pz	<.001* BF=5.4 ++	.506 BF=-3.17 - -	<.001* BF=5.98 ++	.811 BF=-3.74 - -
3	*Location for Unexpected*	Unexpected Diff Fz - Unexpected Diff Pz	<.001* BF=13.19 ++	<.001* BF=20.2 +++	<.001* BF=15.79 ++	<.001* BF=15.61 ++
4	*Location for Expected*	Expected Diff Fz - Expected Diff Pz	<.001* BF=7.67 ++	<.001* BF=12 ++	<.001* BF=10.42 ++	<.001* BF=12.26 ++
5	*Expectancy for Reward*	Unexpected Reward Fz- Expected Reward Fz	<.001* BF=4.64 ++ [9]	<.001* BF=11.83 ++	0.442 BF=-2.87 -	<.001* BF=6.79 ++
6	*Expectancy for NoReward*	Unexpected NoReward Fz - Expected NoReward Fz	0.369 BF=-2.92 -	.023* BF=-0.42	<.001* BF=3.45 ++ [10]	.696 BF=-3.8 - -
7	*Valence for Unexpected*	Unexpected Reward Fz- Unexpected NoReward Fz	<.001* BF=38.94 +++	<.001* BF=41.94 +++	<.001* BF=24.87 +++	<.001* BF=26.36 +++
8	*Valence for Expected*	Expected Reward Fz- Expected NoReward Fz	<.001* BF=23.99 +++	<.001* BF=33.6 +++	<.001* BF=26.37 +++	<.001* BF=25.21 +++

Note. First value refers to the p-Map value, an asterisk indicating a significant effect, BF = BF_{10} = logarithmic Bayes Factor *(BF)* of H1. +++/- - - indicates strong evidence in favor of/against H1. ++/- - positive evidence. +/- weak evidence. Diff = Difference NoReward – Reward outcome.

[9] The pre-registered quantification showed an opposite effect: .092; BF=-0.94.
[10] The pre-registered quantification showed an opposite effect: .239; BF=-2.34.

When using the original pre-registered quantification time window following the original study, similar results were obtained unless specified otherwise in the footnotes. RobTest = Robustness Test.

Table 4: *Bayes Factor Analysis for the different Robustness Tests of P300 component at Pz*

			Original		Current Standards	
			Peak	Mean	Peak	Mean
1	*Expectancy for Reward*	Unexpected Reward - Expected Reward	<.001* BF=28.76 +++	<.001* BF=26.83 +++	<.001* BF=27.88 +++	<.001* BF=23.07 +++
2	*Expectancy for NoReward*	Unexpected NoReward - Expected NoReward	<.001* BF=16.41 ++	<.001* BF=17.94 ++	<.001* BF=21.49 +++	<.001* BF=14.84 ++
3	*Valence for Unexpected*	Unexpected Reward - Unexpected NoReward	<.001* BF=8.85 ++	<.001* BF=7.84 ++	<.001* BF=7.84 ++	.001* BF=5.24 ++
4	*Valence for Expected*	Expected Reward - Expected NoReward	<.001* BF=8.07 ++	<.001* BF=4.71 ++	<.001* BF=7.31 ++	<.001* BF=4.72 ++

Note. First value refers to the p-Map value, an asterisk indicating a significant effect, BF = BF_{10} = logarithmic Bayes Factor *(BF)* of H1. +++ indicates strong evidence in favor of H1. ++ positive evidence.

3.4. Meta-Analysis (Robustness test 5)

For the meta-analysis, forest and funnel plots were computed. We report and plot median and distribution of the weighted effect sizes, 95% confidence intervals, and the number of labs successfully replicating the original effect.

The meta-analysis on the FRN/RewP showed significant main effects of Expectancy (r = 0.32, p < .001, 95% CI [0.22, 0.42], $Q(12)$ = 2.3, p = .999, I^2 = 0.0%) and Location (r = 0.60, p < .001, 95% CI [0.52, 0.66], $Q(12)$ = 8.6, p =.733, I^2 = 0.0%), as well as interaction between these two factors (r = 0.16, p = .005, 95% CI [0.05, 0.27], $Q(12)$ = 13.1, p = .359, I^2 = 6.7%). The large main effect of Location was robustly detected across all labs except one (i.e., 12 out of 13 labs showed a significant effect in the expected direction, with the FRN/RewP the largest at Fz > Cz > Pz). While all labs showed that the FRN/RewP was numerically larger for unexpected compared to expected outcomes, this relatively small effect was only significant in a few of them(i.e., 4 out of 13 labs showed it). Moreover, the interaction between Location and Expectancy was only significant in 2 out of 13 labs, and some of them showed even opposite effects (see Figure 5 and Supplementary Figure 4).

The meta-analysis on the P300 showed significant main effects of Expectancy (r = 0.63, p < .001, 95% CI [0.56, 0.69], $Q(12)$ = 9.5, p = .661, I^2 = 0.0%) and Valence (r = 0.59, p < .001, 95% CI [0.49, 0.68], $Q(12)$ = 20.3, p = .062, I^2 = 41.3%), while the interaction between them was not significant (r = 0.07, p = .23, 95% CI [-0.04, 0.17], $Q(12)$ = 11.9, p = .457, I^2 = 0.0%). The large main effect of Expectancy was robustly detected across all labs (i.e., all 13 labs showed a

significant effect in the expected direction, with the P300 being larger for unexpected compared to expected outcomes). Similarly, the large main effect of Valence was robustly detected in a majority of labs (i.e., 11 out of 13 labs showed a significant effect in the expected direction, with the P300 being larger for reward compared to no-reward outcomes). In comparison, the interaction between Valence and Expectancy was only significant in one lab, where the effect was reversed compared to most other labs (see Figure 5 and Supplementary Figure 4).

For all effects, the aggregated effect sizes across all labs fell within the estimated confidence interval of the original sample, which were quite wide. However, for the previously reported significant effects (i.e., main effect of Location for the FRN, main effect of Expectancy for the P300 component), the aggregated effect sizes were smaller than the ones reported in the original study. In comparison, for the previously reported non-significant effects (i.e., main effect of Expectancy for the FRN, main effect of Valence for the P300 component), the aggregated effect sizes were larger than the ones reported in that study.

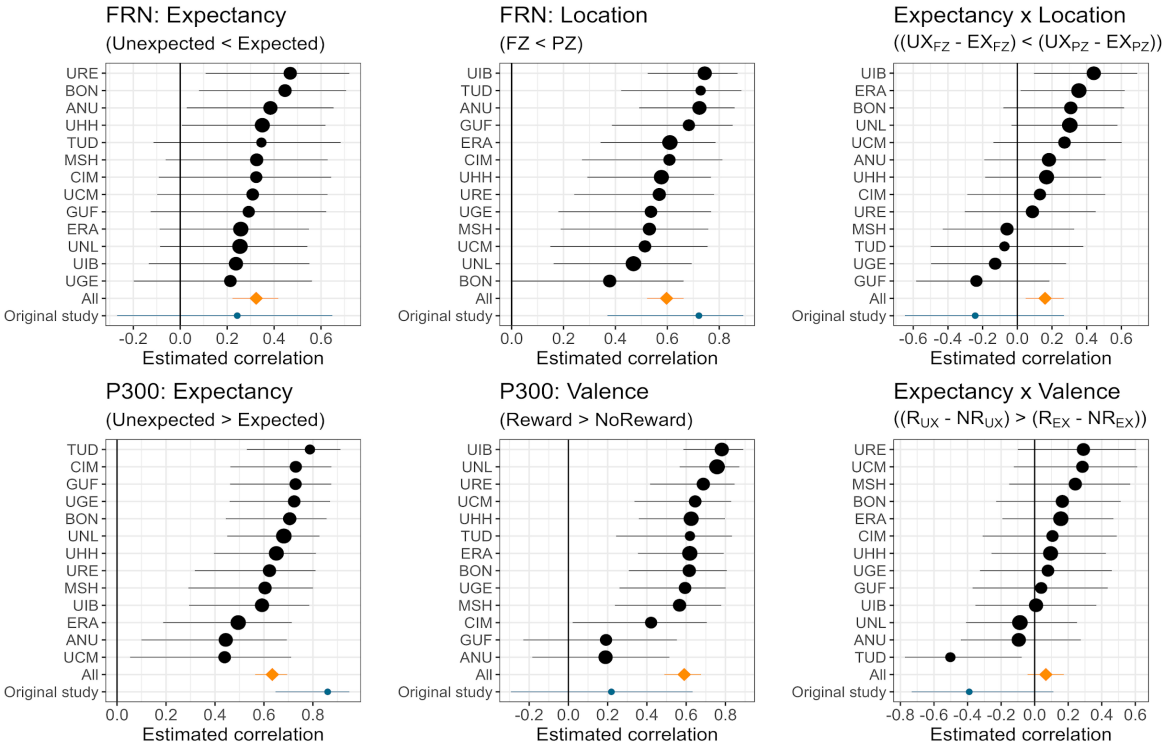


Figure 5. Forest Plots. Correlation coefficients (converted from partial eta squared) for various laboratories. Circle size corresponds to sample size, indicating the robustness of findings in each lab. The orange square shows the meta-analytically aggregated score. The blue circle shows the effect size from Hajcak et al. 2003 (derived from the reported F statistic). Correlation coefficients are coded in such a way that positive values are evidence in favor of the expected effect under consideration (noted in the caption). UX = Unexpected. EX = Expected. R = Reward. NR = NoReward. Please note that the FRN is a negative potential, hence a smaller (more negative) amplitude shows a stronger effect. ANU = Australian National University, Australia. BON = Bond University, Australia. CIM = Central Institute of Mental Health Mannheim, Germany. ERA = Erasmus University Rotterdam, The Netherlands. GUF = Goethe University Frankfurt am Main, Germany. MSH = Medical School Hamburg, Germany. TUD = Technical University Dresden, Germany. UCM = CINPSI Neurocog UCMaule, Chile. UGE = Ghent

University, Belgium. UHH = University Hamburg, Germany. UIB = University of Bergen, Norway. UNL = Leiden University, The Netherlands. URE = University of Regensburg, Germany

3.5. Temporospatial Principal Component Analysis (Robustness Test 6)

For the PCA analysis, the data of 230 participants coming from 8 labs could be used (the other ones did not include the relevant channels). Based on the time course and scalp distribution, four temporospatial factors were identified that closely corresponded to the FRN/RewP and P300 components. One of these factors captured the spatiotemporal variations of the FRN/RewP component, while the remaining ones captured that of the P300 component (see Figure 6).

The PCA factor TF03SF1, corresponding to the FRN/RewP component, exhibited a peak latency at 276 ms over the central area (maximal at Cz). The robust ANOVA revealed a significant main effect of Valence ($T_{WJt}/c_{1.0,198.0}$ = 271.02, p < .001, MSe = 38.75), exhibiting a larger positivity for reward than no-reward outcomes (M_{Reward} = 12.08 µV, sd = .04 vs. M_{NoReward} = 6.15 µV, SD = .03). In contrast, the main effect of Expectancy was not significant ($T_{WJt}/c_{2.0,176.0}$ = 1.67, p = .189, MSe = 13.01). Moreover, the interaction between them was significant ($T_{WJt}/c_{2.0,176.0}$ = 6.16, p < .002, MSe = 5.28), with this PCA factor differentiating better reward from no-reward outcomes for unexpected compared to expected outcomes. The positivity was larger for unexpected reward compared to expected rewards, while the opposite pattern was true for no-reward outcomes ($M_{\text{Unexpected Reward}}$ = 12.38 µV, sd = .04 vs. $M_{\text{Expected Reward}}$ = 11.66 µV, sd = .03 vs. $M_{\text{Unexpected NoReward}}$ = 6.29 µV, sd = .03 vs. $M_{\text{Expected NoReward}}$ = 6.35 µV, sd = .03).

The PCA factor TF04SF1, which corresponded to the P300 component, exhibited a peak latency at 366 ms over the central area (maximal at Cz). Although the robust ANOVA revealed no significant main effect of Valence ($T_{WJt}/c_{1.0,198.0}$ = 1.03, p = .311, MSe = 15.76), the effect of Expectancy reached significance ($T_{WJt}/c_{2.0,176.0}$ = 37.78, p < .001, MSe = 7.91), exhibiting the largest positivity for the unexpected outcomes ($M_{\text{Unexpected}}$ = 7.50 µV, sd = .02 vs. M_{Expected} = 5.85 µV, sd = .02 vs. M_{Neutral} = 6.22 µV, sd = .02). The interaction between Valence and Expectancy was not significant ($T_{WJt}/c_{2.0,176.0}$ = 3.29, p = .0389, MSe = 4.98).

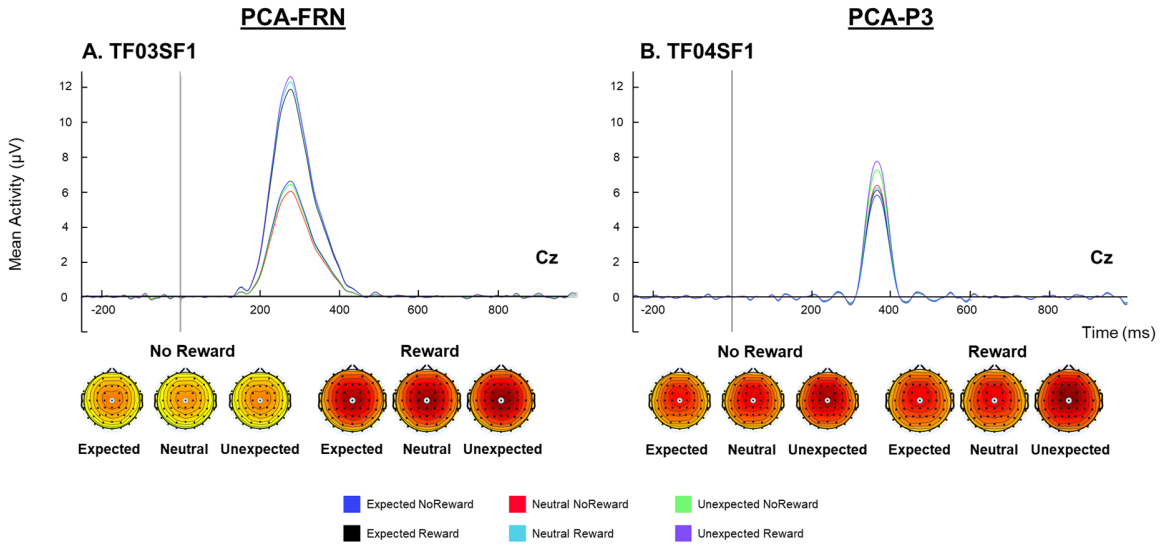


Figure 6. Activation over time and topographical plots for the PCA factors corresponding to the RewP and P300 components: (A) The factor TF03SF1 (corresponding to the RewP) peaks at 276 ms at the central area. (B) The factor TF04SF1 (corresponding to the P3) peaks at 366 ms at the central area.

Two additional PCA factors could be related to the P300 component and are described in the supplementary material since their latency was later than the average peak of this ERP component (after 400 ms, although still falling within the time interval of the P300 component according to some models; see (Polich, 2007). These two additional factors were both significantly modulated by Valence, while only one of them additionally showed a significant main effect of Expectancy. None of them showed a significant interaction effect, see Supplementary section 4.

4. Discussion

In this study, we directly replicated Hajcak et al. (2005) as part of the #EEGManyLabs project (Pavlov et al., 2021). We examined the sensitivity of the FRN and P300 components to outcome valence and expectancy using a simple guessing task. Hajcak et al. (2005) found that the FRN distinguished reward from no-reward outcomes regardless of expectancy, while the P300 differentiated unexpected from expected outcomes, independent of valence. This led to a two-stage model of feedback processing: valence is processed at the FRN level, while expectancy mostly influences the P300. Our replication, with an unprecedented sample size of up to 360 participants across 13 laboratories worldwide, partly corroborates these findings but contradicts this simple two-stage model. Unlike Hajcak et al. (2005), we found that both the FRN/RewP and

the P300 components were significantly modulated by both outcome expectancy and valence. In addition to the exact replication using the same EEG pre-processing and scoring methods, we conducted several robustness tests, a meta-analysis including laboratory as a variable, and a PCA. These methods consistently confirmed our findings for the FRN/RewP and P300 components.

The original study reported significant effects of expectancy only for the P300 (valence for the FRN was not tested) while, for the FRN/RewP and P300 components, it reported null-effects of expectancy and valence, respectively. Therefore, we aimed for a large sample size in our study to detect small but relevant effects (Paul et al., 2020). In comparison, the original study had a modest sample size (n=17), common in neurophysiology at that time (Picton et al., 2000). However, with a well-powered sample, we observed that expectancy had a small to moderate effect on the FRN/RewP component. With only 17 participants, detecting a similar significant effect would have been rather unlikely given that the statistical power to detect an effect of $\eta p^2 = 0.08$ with 17 participants is only around 40%. Consequently, the previously reported "insensitivity" of the FRN/RewP to expectancy was most likely a false negative finding, emphasizing that absence of evidence does not equate to evidence of absence.

Our study shows instead that the FRN/RewP is robustly modulated by expectancy, albeit to a lesser extent than by valence, and to a lesser extent than the P300 component. This result challenges the view that the FRN/RewP solely represents binary outcome valence processing (Hajcak et al., 2006; Kujawa et al., 2013). To explain it, the reinforcement learning framework provides a more plausible model, according to which the FRN/RewP captures activity in a dopaminergic fronto-striatal network where both valence and expectancy are processed concurrently (Holroyd & Coles, 2002; Ullsperger, Fischer, et al., 2014). At the same time, it remains to be determined which role the P300 component could play in this ERN-RL framework. Moreover, using a guessing task rather than a learning task, our replication indicates that reinforcement learning (see Sutton & Barto, 1998) is not required to produce these ERP effects. This implies that the cue information about reward probability was sufficient to influence feedback processing. These findings support the idea that subjective expectancy, rather than reward probability maximization, could actually drive these FRN/RewP amplitude changes (Walentowska et al., 2019). Notably, other ERP findings suggest that even when reward probabilities were held constant, the FRN/RewP amplitudes could vary depending if reward probabilities were perceived as better or worse than previously experienced (Mushtaq et al., 2013). In this vein, later results from Hajcak et al. (2007) are also informative: using the same guessing task as used here, the authors asked participants about their reward expectations either before or after the information cue. They found that the FRN/RewP component was sensitive to the expectancy manipulation only when participants rated their expectations after the presentation of the information cue, suggesting that this effect depends on the close coupling of (subjective) predictions and outcomes. In the current study, our results show that participants reported paying attention to both the reward probability cue and the feedback, possibly indicating they sought to maximize reward, even though outcomes were unrelated to any behavioral strategy. Thus, it is possible that the effect of (objective) expectancy on the FRN/RewP component becomes larger the more explicitly subjective expectations align with manipulated variables. We suggest that a potentially fruitful line of future study could be to directly compare the impact of subjective and objective reward probabilities.

The second discrepancy worth-mentioning between our results and the original study is that also the P300 component is robustly modulated by both valence and expectancy, and not only expectancy as postulated by the two-stage model outlined above. While expectancy's influence on the P300 is well-documented across various domains (Polich, 2007), valence effects on this ERP component during performance monitoring are less consistent (Ullsperger, Fischer, et al., 2014). Some have even argued that it is blind to outcome valence (Hajcak et al., 2006; Kujawa et al., 2013). Our replication clearly demonstrates that the P300 amplitude is significantly modulated by outcome valence, being larger for reward than no-reward outcomes. This suggests that its amplitude variations likely reflect a motivational effect (Nieuwenhuis et al., 2005; San Martín, 2012). Although the specific processes underlying the P300 remain unclear (Verleger, 2020), our findings indicate that this component is enhanced for favorable outcomes, possibly reflecting approach motivation (Harmon-Jones et al., 2013). This aligns with a study on social feedback processing, which also found enhanced P300 activity for favorable, expected outcomes (van der Veen et al. 2014). Nevertheless, since we did not include a loss condition (only no-reward vs. reward), it remains an open question how these effects compare to unfavorable outcomes. Additional EEG research is needed to address this question and directly assess the extent to which motivationally relevant or meaningful outcomes could influence the P300 component (Glazer et al., 2018; San Martín, 2012; Stewardson & Sambrook, 2020). In this context, it appears important to clarify whether relevance, memory updating, or perhaps another cognitive or emotional process drives this neurophysiological effect.

Given that the FRN/RewP and P300 components rapidly follow each other, overlapping effects of expectancy and valence may be artificially inflated. This makes our additional PCA analysis particularly important, as both components clearly distinguish between reward and no-reward outcomes. The PCA allowed us to disentangle successive and overlapping ERP components (Dien, 2012). While carefully controlling for the influence of other spatiotemporal components, the PCA revealed that the valence effect at the FRN/RewP level was distinct and independent from that of the P300. Our findings therefore suggest that valence processing is multifaceted and influences both the FRN/RewP and P300, which likely capture distinct facets of it. Speculatively, the FRN/RewP may reflect early hedonic feedback processing ("liking"), while the subsequent P300 may represent its motivational value ("wanting"), consistent with theoretical frameworks that decompose brain pleasure mechanisms into liking and wanting components (see Berridge et al., 2012). A related effect could be shown when considering saturation to (e.g. food-related) rewards, which affected only the P300 component (Huverman et al. 2021). Even more speculatively, a similar division might be applied to the processing of outcome expectancy because the PCA analysis confirmed distinct and independent effects of it on the FRN/RewP and the P300. This implies that, similar to valence, expectancy processing during performance monitoring could involve multiple components. While the PCA effectively disentangles these components, the functional significance of these successive expectancy (as well as valence) effects remains challenging to grasp. Because it could not be addressed directly with the current ERP analyses, future studies are needed to shed light on it and eventually improve or amend current theoretical models of performance monitoring. Moreover, at the methodological level, since we used visual inspection to select the main PCA factors corresponding to the FRN/RewP

and P300, we believe that replicability could be enhanced in the future if automated procedures or algorithms would be used to carry out this selection.

Based on our results, one could hypothesize that the FRN/RewP reflects a "crude" reward prediction error in a midbrain-dependent fronto-striatal loop (Schultz, 2016). Consistent with this hypothesis, single-trial ERP studies have shown that both FRN/RewP and P300 are influenced by prediction errors but this influence varies depending on the context (Hoy et al. 2021, Weber & Bellebaum, 2024). Interestingly, even cerebellar output is crucial for learning from action outcomes, as disruptions in cerebellar function impair the FRN/RewP component (Huvermann et al., 2024). The fronto-striatal reward prediction error signal is then being relayed to areas such as the hippocampus or entorhinal cortex involved in memory or reinforcement learning, potentially giving rise to the P300 component (Soltani & Knight, 2000).

Besides the theoretical implications and better functional delineation of the FRN/RewP and P300 components during performance monitoring, our replication highlights their sensitivity to different EEG data processing methods. Embedded in the #EEGManyLabs project, our replication aimed to address methodological limitations of previous EEG research, such as small sample sizes and lack of preregistration (Pavlov et al., 2021). We performed an almost exact replication with sufficient statistical power and supplemented it with robustness tests, including a PCA and a meta-analysis. Overall, these analyses largely concurred on a robust amplitude modulation of the FRN/RewP by expectancy. Nonetheless, there are some differences between them worth mentioning, as they might explain some of the discrepant results reported earlier in the literature. First, the peak-scoring method showed the expectancy effect of the FRN/RewP most robustly (see also Paul et al. 2019). In contrast, the mean-scoring method yielded more topographical precision as it was confined to Fz, where this component is expected to reach its maximum amplitude given its intracranial generators are presumably located in the dorsal medial prefrontal cortex (Hauser et al. 2014). At the same time, the liability of the peak-scoring to noise (see Luck, 2005) led to a larger SME as a measure of within-subject variability across trials (see supplementary Table 1). Fortunately, the pre-processing strategies did not have a large influence on the pattern of results. However, they were aimed to be as similar as possible, with the largest difference concerning the correction of ocular artifacts. Other important methodological choices, e.g., the choice of reference or the time-window used to define the ERP components, were not investigated, but are probably worth exploring further in future EEG studies. Multiverse analyses, which systematically explore the influence of methodological choices across multiple analytical pipelines, could also provide valuable insights into these questions (see Clayson, 2024).

Aligned with the ERN-RL framework, the FRN/RewP difference was found to be larger for unexpected events, with a stronger response observed for unexpected versus expected outcomes. Alongside the traditional difference-wave approach (reward versus no-reward) used to assess the influence of expectancy effects on the FRN/RewP component, we further assessed the components separately for reward and no-reward outcomes to investigate whether the expectancy effect was driven by one type of outcome, providing additional insights into its mechanism. Prior research suggests that the RewP (in response to reward) can serve as the counterpart to the FRN (in response no-reward or loss) with opposite polarity (Proudfit, 2015, Kappenman et al., 2021). Although difference-waves are used to analyze these effects in this framework, the RewP and FRN's different spatio-temporal properties may obscure distinct modulatory effects of expectancy and valence for reward and no-reward outcome, respectively

(Gheza et al., 2018). Our findings strongly support this distinction, showing that expectancy effects were stronger and more robust for rewards than no-rewards, indicating that the RewP/FRN component was boosted in particular in response to unexpected reward outcomes, while the RewP/FRN component was not influenced by the expectancy of no-reward. Bayes factors provided evidence for the absence of an expectancy effect for no-rewards (when using mean-scoring, which is more suitable to define the RewP in the absence of a clear peak). This finding is not surprising, as rewards were more relevant to participants in the current task, while no-rewards were presumably less informative. Therefore, the relevance or informativeness of feedback, which may be closely linked to the participants' curiosity and motivation for information-seeking, should be considered in future study designs (Kidd & Hayden 2014). It is possible that expectancy impacts feedback processing at the level of the FRN/RewP component only when the feedback is meaningful to the participant (Walentowska et al., 2018). Additionally, prior research using the absence of aversive outcomes as positive outcomes (i.e., "rewarding") has often failed to find that the RewP/FRN component (defined as a difference score) is larger for unexpected than expected feedback, highlighting the importance of outcome type in determining expectancy effects (e.g., Talmi et al., 2013; Bauer et al., 2024). More broadly, these findings suggest that using difference waves may not be ideal for examining the modulatory effects of expectancy on early performance monitoring ERP components, as this method can obscure potentially asymmetrical effects on the overlapping RewP and FRN components.

At the methodological level, our series of robustness tests allowed us to compare various analytical approaches for data collected across multiple labs. These approaches included an ANOVA on the entire dataset without accounting for potential differences between labs, a Bayesian multilevel model (MLM) with random intercepts and slopes for each lab, and a random-effects meta-analysis across all 13 labs, which accounted for the lab effect and estimated heterogeneity. Importantly, accounting for the differences between labs did not significantly alter the effect sizes (e.g., η_p^2 = 0.08 in ANOVA vs. η_p^2 = 0.10 in the meta-analysis for the FRN expectancy effect, and η_p^2 = 0.32 in ANOVA vs. η_p^2 = 0.35 in the meta-analysis for the P300 valence effect). The use of Bayesian MLM, compared to the meta-analysis, did not noticeably affect the results either, at least for the peak scoring approach.

We collected data from 13 laboratories across seven countries on three continents. Despite noticeable differences in hardware and potential variations in local populations, the overall effect, exemplified by the expectancy effect on FRN, remained consistent. Strikingly, conventional heterogeneity estimates indicated no variability. This result is important because it indirectly suggests that the effects in this task are quite robust. Notably, the effect size for the expectancy effect on FRN in the ANOVA in our replication turned out to be exactly the same as our estimate of the effect size in the original study. Moreover, our effect sizes for all effects of interest fell within the confidence interval of the original study. The diverse nature of our sample, along with the absence of variability in the results, further supports the robustness of the observed effects.

In addition to these methodological insights, our study provides some recommendations for future EEG studies on the FRN/RewP and P300 components regarding sample size estimation, should the same task be used (see supplementary Table 8). In short, for each ERP component and effect under consideration (i.e., Location, Valence, Expectancy, or their interaction), we have used the effect size reported in this study and computed a sample size estimation. We believe this information could be valuable to researchers working on performance monitoring. Moreover,

because the data and scripts of this replication are publicly available, they could easily be used in future studies to perform additional analyses (e.g., time-frequency decompositions). Similarly, our data could be pooled together with other EEG data sets available in the literature and contribute to mega-studies or mega-analyses (Costafreda, 2009). These efforts would have the potential to provide a more precise estimate of the effect size under scrutiny or to identify possible moderators (e.g., learning, different feedback types or stimuli used). Additionally, since we collected some personality questionnaires, pooling with existing data could allow the investigation of interindividual differences in feedback processing. Furthermore, this study highlights the broader significance of replication studies in advancing psychological theories. Replication not only validates previous findings but also refines and challenges existing theories in cognitive neuroscience, ensuring that they are robust and generalizable. Thus, we can uncover nuances and inconsistencies that lead to a deeper understanding of psychological processes.

In conclusion, our replication underscores the complexity of feedback processing in the brain and reveals several advantages of a large and collaborative EEG data collection to gain novel insights. Crucially, we found no support of the two-stage model of feedback processing. Instead, our new results suggest that the premises of the ERN-RL model might also include the P300 component, besides the FRN/RewP. In light of them, we suggest an integrated model of evaluative feedback processing where both valence and expectancy are concurrently processed across multiple stages. Furthermore, we advocate for more stringent methods, including the use of preregistration and the consideration of effect sizes to determine appropriate sample sizes, and hope the present replication and associated resources could be used to guide future research on the electrophysiological correlates of feedback processing.

5. Declaration of Interest

The authors declare that there is no conflict of interest. Funders and employers had no role in study design or the decision to submit the work for publication.

6. Acknowledgements

The #EEGManyLabs project is funded by a DFG grant (PA 4005/1-1) provided to YGP and a UK Research and Innovation Biotechnology and Biological Sciences Research Council award (BB/X008428/1) to FM. B.N.J. was supported by Australian Research Council (DE220100739); F.B. was supported by DFG (BU 3255/1-2); K.P. was supported by DFG (PA 4014/2-2); J.P. was supported by DFG (PE 2077/6-1; PE 2077/7-1); Y.L.S. was supported by European Union (ERC-2018-StG-PIVOTAL-758898); D.M.P. was supported by South-Eastern Norway Regional Health Authority (2021046) in 2022.

7. References

Åkerstedt, T., & Gillberg, M. (1990). Subjective and Objective Sleepiness in the Active Individual. *International Journal of Neuroscience*, *52*(1–2), 29–37. https://doi.org/10.3109/00207459008994241

Bauer, E. A., Watanabe, B. K., & MacNamara, A. (n.d.). Reinforcement learning and the reward positivity with aversive outcomes. *Psychophysiology*, 61(4), e14460. https://doi.org/10.1111/psyp.14460

Bernat, E. M., Nelson, L. D., & Baskin-Sommers, A. R. (2015). Time-frequency theta and delta measures index separable components of feedback processing in a gambling task. *Psychophysiology*, *52*(5), 626–637. https://doi.org/10.1111/psyp.12390

Berridge, C. W., Schmeichel, B. E., & España, R. A. (2012). Noradrenergic modulation of wakefulness/arousal. *Sleep Medicine Reviews*, *16*(2), 187–197. https://doi.org/10.1016/j.smrv.2011.12.003

Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R Journal*, *10*(1), 395–411. https://doi.org/10.32614/rj-2018-017

Campbell, J. I. D., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods*, *44*(4), 1255–1265. https://doi.org/10.3758/s13428-012-0186-0

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. In *Journal of Personality and Social Psychology* (Vol. 67, pp. 319–333). https://doi.org/10.1037/0022-3514.67.2.319

Chatrian, G. E., Lettich, E., & Nelson, P. L. (1985). Ten percent electrode system for topographic studies of spontaneous and evoked EEG activities. *American Journal of EEG Technology*, *25*(2), 83–92.

Clayson, P. E. (2024). Beyond single paradigms, pipelines, and outcomes: Embracing multiverse analyses in psychophysiology. *International Journal of Psychophysiology*, 112311. https://doi.org/10.1016/j.ijpsycho.2024.112311

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. *Journal of Neuroscience Methods*.

Costafreda, S. G. (2009). Pooling fMRI data: Meta-analysis, mega-analysis and multi-center studies. *Frontiers in Neuroinformatics*, *3*(SEP), 1–8. https://doi.org/10.3389/neuro.11.033.2009

Courchesne, E., Hillyard, S. A., & Courchesne, R. Y. (1977). P3 Waves to the Discrimination of Targets in Homogeneous and Heterogeneous Stimulus Sequences. *Psychophysiology*, *14*(6), 590–597. https://doi.org/10.1111/j.1469-8986.1977.tb01206.x

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dickey, J. M., & Lientz, B. P. (1970). The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226. https://doi.org/10.1214/aoms/1177697203

Dien, J. (2010a). Evaluating two-step PCA of ERP data with Geomin, Infomax, Oblimin, Promax, and Varimax rotations. *Psychophysiology*, *47*(1), 170–183. https://doi.org/10.1111/j.1469-8986.2009.00885.x

Dien, J. (2010b). The ERP PCA Toolkit: An open source program for advanced statistical analysis of event-related potential data. *Journal of Neuroscience Methods*, *187*(1), 138–145. https://doi.org/10.1016/j.jneumeth.2009.12.009

Dien, J. (2012). Applying Principal Components Analysis to Event-Related Potentials: A Tutorial. *Developmental Neuropsychology*, *37*(6), 497–517. https://doi.org/10.1080/87565641.2012.697503

Ethridge, P., & Weinberg, A. (2018). Psychometric properties of neural responses to monetary and social rewards across development. *International Journal of Psychophysiology*, *132*(January), 311–322. https://doi.org/10.1016/j.ijpsycho.2018.01.011

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian Inference for Psychology. *Psychonomic Bulletin and Review*, *25*(1), 5–34. https://doi.org/10.3758/s13423-017-1262-3

Ferdinand, N. K., Mecklinger, A., Kray, J., & Gehring, W. J. (2012). The Processing of Unexpected Positive Response Outcomes in the Mediofrontal Cortex. *Journal of Neuroscience*, *32*(35), 12087–12092. https://doi.org/10.1523/JNEUROSCI.1410-12.2012

Foti, D., Weinberg, A., Dien, J., & Hajcak, G. (2011). Event-related potential activity in the basal ganglia differentiates rewards from nonrewards: Response to commentary. *Human Brain Mapping*, *32*(12), 2267–2269. https://doi.org/10.1002/hbm.21357

Gable, P. A., Paul, K., Pourtois, G., & Burgdorf, J. (2021). Utilizing electroencephalography (EEG) to investigate positive affect. *Current Opinion in Behavioral Sciences*, *39*, 190–195. https://doi.org/10.1016/j.cobeha.2021.03.018

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (2018). The Error-

Related Negativity. *Perspectives on Psychological Science*, *13*(2), 200–204. https://doi.org/10.1177/1745691617715310

Gehring, W. J., & Willoughby, A. R. (2002). The Medial Frontal Cortex and the Rapid Processing of Monetary Gains and Losses. *Science*, *295*(5563), 2279–2282. https://doi.org/10.1126/science.1066893

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian Regression Models. *American Statistician*, *73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Gerlitz, J.-Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *Research Notes 4*, *May*, 1–44. https://doi.org/10.1016/j.jsis.2005.07.003

Gheza, D., Paul, K., & Pourtois, G. (2018). Dissociable effects of reward and expectancy during evaluative feedback processing revealed by topographic ERP mapping analysis. *International Journal of Psychophysiology*, *132*(November), 213–225. https://doi.org/10.1016/j.ijpsycho.2017.11.013

Glazer, J. E., Kelley, N. J., Pornpattananangkul, N., Mittal, V. A., & Nusslock, R. (2018). Beyond the FRN: Broadening the time-course of EEG and ERP components implicated in reward processing. *International Journal of Psychophysiology*, *132*(2), 184–202. https://doi.org/10.1016/j.ijpsycho.2018.02.002

Gratton, G., Coles, M. G. . H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*(4), 468–484. https://doi.org/10.1016/0013-4694(83)90135-9

Gu, Y., Liu, T., Zhang, X., Long, Q., Hu, N., Zhang, Y., & Chen, A. (2021). The Event-Related Potentials Responding to Outcome Valence and Expectancy Violation during Feedback Processing. *Cerebral Cortex*, *31*(2), 1060–1076. https://doi.org/10.1093/cercor/bhaa274

Guthrie, E. R. (1942). Conditioning: A theory of learning in terms of stimulus, response, and association. *Teachers College Record*, *43*(10), 17–60.

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

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*(2), 148–154. https://doi.org/10.1016/j.biopsycho.2005.04.001

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology*, *44*(6), 905–912. https://doi.org/10.1111/j.1469-8986.2007.00567.x

HajiHosseini, A., Rodríguez-Fornells, A., & Marco-Pallarés, J. (2012). The role of beta-gamma oscillations in unexpected rewards processing. *NeuroImage*, *60*(3), 1678–1685. https://doi.org/10.1016/j.neuroimage.2012.01.125

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: new insights into the localization,

meaning and network organization. *Neuroimage*, 84, 159-168. https://doi.org/10.1016/j.neuroimage.2013.08.028

Herrmann, C. S., & Knight, R. T. (2001). Mechanisms of human attention: event-related potentials and oscillations. *Neuroscience & Biobehavioral Reviews*, *25*(6), 465–476. https://doi.org/10.1016/S0149-7634(01)00027-6

Hoffman. (2014). The No-U-Turn Sample. *Journal of Machine Learning Research*, *15*, 1593–1623. http://mcmc-jags.sourceforge.net

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Holroyd, C. B., Krigolson, O. E., Baker, R., Lee, S., & Gibson, J. (2009). When is an error not a prediction error ? An electrophysiological investigation. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(1), 59–70. https://doi.org/10.3758/CABN.9.1.59

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Cohen, J. D., Nieuwenhuis, C. A. S., Nick, Y., & 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

Hoy, C. W., Steiner, S. C., & Knight, R. T. (2021). Single-trial modeling separates multiple overlapping prediction errors during reward processing in human EEG. *Communications Biology*, *4*(1), 1–17. https://doi.org/10.1038/s42003-021-02426-1

Huvermann, D. M., Bellebaum, C., & Peterburs, J. (2021). Selective Devaluation Affects the Processing of Preferred Rewards. *Cognitive, Affective and Behavioral Neuroscience*, *21*(5), 1010–1025. https://doi.org/10.3758/s13415-021-00904-x

Huvermann, D. M., Berlijn, A. M., Thieme, A. G., Erdlenbruch, F., Groiss, S. J., Deistung, A., … Peterburs, J. (2024, May 22). The cerebellum contributes to prediction error coding in reinforcement learning - complementary evidence from stroke patients and from cerebellar transcranial magnetic stimulation. *Preprint*. https://doi.org/10.31219/osf.io/a8hbx

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Johnson, R., & Donchin, E. (1980). P300 and Stimulus Categorization: Two Plus One is not so Different from One Plus One. *Psychophysiology*, *17*(2), 167–178. https://doi.org/10.1111/j.1469-8986.1980.tb00131.x

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. *NeuroImage*, *225*(October 2020), 117465. https://doi.org/10.1016/j.neuroimage.2020.117465

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, 88(3), 449-460. doi: 10.1016/j.neuron.2015.09.010. PMID: 26539887; PMCID: PMC4635443.

Klawohn, J., Meyer, A., Weinberg, A., & Hajcak, G. (2020). Methodological choices in event-related potential (ERP) research and their impact on internal consistency reliability and individual differences: An examination of the error-related negativity (ERN) and anxiety. *Journal of Abnormal Psychology*, *129*(1), 29–37. https://doi.org/10.1037/abn0000458

Krigolson, O. E. (2018). Event-related brain potentials and the study of reward processing: Methodological considerations. *International Journal of Psychophysiology*, *132*(November

2017), 0–1. https://doi.org/10.1016/j.ijpsycho.2017.11.007

Kujawa, A., Smith, E., Luhmann, C., & Hajcak, G. (2013). The feedback negativity reflects favorable compared to nonfavorable outcomes based on global, not local, alternatives. *Psychophysiology*, *50*(2), 134–138. https://doi.org/10.1111/psyp.12002

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(NOV), 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Luck, S. J. (2005). *An introduction to the event related potential technique* (M. S. Gazzaniga (ed.)). MIT Press.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology*, *58*(6), 1–15. https://doi.org/10.1111/psyp.13793

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*, *52*(2), 464–488. https://doi.org/10.3758/s13428-019-01246-w

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, *10*(December), 1–14. https://doi.org/10.3389/fpsyg.2019.02767

Mills, J. (2018). *Objective Bayesian Precise Hypothesis Testing*. https://doi.org/10.13140/RG.2.2.13158.32328

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-Related Brain Potentials Following Incorrect Feedback in a Time-Estimation Task: Evidence for a “Generic” Neural System for Error Detection. *Journal of Cognitive Neuroscience*, *9*(6), 788–798. https://doi.org/10.1162/jocn.1997.9.6.788

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

Mushtaq, F., Stoet, G., Bland, A. R., & Schaefer, A. (2013). Relative changes from prior reward contingencies can constrain brain correlates of outcome monitoring. *PLoS One*, 8(6), e66350. https://doi.org/10.1371/journal.pone.0066350

Nalborczyk, L., Batailler, C., Loevenbruck, H., Vilain, A., & Bürkner, P. C. (2019). An introduction to bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006

Natarajan, R., & Kass, R. E. (2000). Reference Bayesian Methods for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *95*(449), 227–237. https://doi.org/10.1080/01621459.2000.10473916

Nieuwenhuis, S. (2004). Sensitivity of Electrophysiological Activity from Medial Frontal Cortex to Utilitarian and Performance Feedback. *Cerebral Cortex*, *14*(7), 741–747. https://doi.org/10.1093/cercor/bhh034

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus-norepinephrine system. *Psychological Bulletin*, *131*(4), 510–532. https://doi.org/10.1037/0033-2909.131.4.510

Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. H. (2004). *Reinforcement-related brain potentials from medial frontal cortex : origins and functional significance*. *28*, 441–448. https://doi.org/10.1016/j.neubiorev.2004.05.003

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Paul, M., Govaart, G. H., & Schettino, A. (2020). Preregistration : A Solution to Undisclosed Analytic Flexibility in ERP Research. *PsyArXiv*, *https://psyarxiv.com/4tgve*.

Paul K, Vassena E, Severo MC, Pourtois G. (2020) Dissociable effects of reward magnitude on fronto-medial theta and FRN during performance monitoring. *Psychophysiology*. Feb;57(2):e13481. doi: 10.1111/psyp.13481. Epub 2019 Oct 2. PMID: 31578739.

Pavlov, Y. G., Adamian, N., Appelhoff, S., Arvaneh, M., Benwell, C. S. Y., Beste, C., Bland, A. R., Bradford, D. E., Bublatzky, F., Busch, N. A., Clayson, P. E., Cruse, D., Czeszumski, A., Dreber, A., Dumas, G., Ehinger, B., Ganis, G., He, X., Hinojosa, J. A., … Mushtaq, F. (2021). #EEGManyLabs: Investigating the replicability of influential EEG experiments. *Cortex*, *144*, 213–229. https://doi.org/10.1016/j.cortex.2021.03.013

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13.

Pfabigan, D. M., Alexopoulos, J., Bauer, H., & Sailer, U. (2011). Manipulation of feedback expectancy and valence induces negative and positive reward prediction error signals manifest in event-related brain potentials. *Psychophysiology*, *48*(5), 656–664. https://doi.org/10.1111/j.1469-8986.2010.01136.x

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. a, Johnson, R., Miller, G. a, Ritter, W., Ruchkin, D. S., Rugg, M. D., & Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology*, *37*(2), 127–152. https://doi.org/10.1111/1469-8986.3720127

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, *52*(4), 449–459. https://doi.org/10.1111/psyp.12370

R-Core-Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Radloff, L.S. (1977). The CES-D Scale: A self-report report depression scale for research in the general population. *Applied Psychological Measuremen*t, 1(3), 385-401. http://dx.doi.org/10.1177/014662167700100306

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, *141*(1), 213–235. https://doi.org/10.1037/bul0000006

Sambrook, T. D., Roser, M., & Goslin, J. (2012). Prospect theory does not describe the feedback-related negativity value function. *Psychophysiology*, *49*, 1533–1544. https://doi.org/10.1111/j.1469-8986.2012.01482.x

San Martín, R. (2012). Event-related potential studies of outcome processing and feedback-guided learning. *Frontiers in Human Neuroscience*, *6*, 1–17. https://doi.org/10.3389/fnhum.2012.00304

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, *18*(1), 23–32. https://doi.org/10.31887/DCNS.2016.18.1/wschultz

Soltani, M., & Knight, R. T. (2000). Neural origins of the P300. *Critical Reviews™ in Neurobiology*, *14*(3–4).

Spencer, K., Dien, J., & Donchin, E. (1999). A componential analysis of the ERP elicited by novel events using a dense electrode array. *Psychophysiology*, *36*(3), 409–414. https://doi.org/10.1017/S0048577299981180

Spencer, K., Dien, J., & Donchin, E. (2001). Spatiotemporal analysis of the late ERP to deviant stimuli. *Psychophysiology*, *38*, 343–358. https://doi.org/10.1111/1469-8986.3820343

Spielberger, C. D., Gorsuch, R., & Lushene, R. (1970). Manual for the State-Trait Anxiety Inventory. In *Education*.

Stewardson, H. J., & Sambrook, T. D. (2020). Evidence for parietal reward prediction errors using great grand average meta-analysis. *International Journal of Psychophysiology*, *152*(April), 81–86. https://doi.org/10.1016/j.ijpsycho.2020.03.002

Sutton, R. S., & Barto, A. G. (1998). *Introduction to Reinforcement Learning*. MIT Press.

Talmi, D., Atkinson, R., & El-Deredy, W. (2013). The feedback-related negativity signals salience prediction errors, not reward prediction errors. *Journal of Neuroscience*, 33(19), 8264-8269. https://doi.org/10.1523/JNEUROSCI.5695-12.2013

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics, 5th ed. In *Using multivariate statistics, 5th ed.* Allyn & Bacon/Pearson Education.

Ullsperger, M., Danielmeier, C., & Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiological Reviews*, *94*(1), 35–79. https://doi.org/10.1152/physrev.00041.2012

Ullsperger, M., Fischer, A. G., Nigbur, R., & Endrass, T. (2014). Neural mechanisms and temporal dynamics of performance monitoring. *Trends in Cognitive Sciences*, *18*(5), 259–267. https://doi.org/10.1016/j.tics.2014.02.009

Van der Veen, F. M., van der Molen, M. W., Sahibdin, P. P., & Franken, I. H. (2014). The heart-break of social rejection versus the brain wave of social acceptance. *Social Cognitive and Affective Neuroscience*, 9(9), 1346-1351. https://doi.org/10.1093/scan/nst120

Verleger, R. (2020). Effects of relevance and response frequency on P3b amplitudes: Review of findings and comparison of hypotheses about the process reflected by P3b. *Psychophysiology*, *57*(7), 1–22. https://doi.org/10.1111/psyp.13542

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Walentowska, W., Moors, A., Paul, K., & Pourtois, G. (2016). Goal relevance influences performance monitoring at the level of the FRN and P3 components. *Psychophysiology*, *53*(7), 1020–1033. https://doi.org/10.1111/psyp.12651

Walentowska, W., Paul, K., Severo, M. C. M. C., Moors, A., & Pourtois, G. (2018). Relevance and uncertainty jointly influence reward anticipation at the level of the SPN ERP component. *International Journal of Psychophysiology*, *132*(November 2017), 287–297. https://doi.org/10.1016/j.ijpsycho.2017.11.005

Walentowska, W., Severo, M. C., Moors, A., & Pourtois, G. (2019). When the outcome is

different than expected: Subjective expectancy shapes reward prediction error at the FRN level. *Psychophysiology*, *56*(12), 1–16. https://doi.org/10.1111/psyp.13456

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews*, *36*(8), 1870–1884. https://doi.org/10.1016/j.neubiorev.2012.05.008

Warren, C. M., & Holroyd, C. B. (2012). The Impact of Deliberative Strategy Dissociates ERP Components Related to Conflict Processing vs. Reinforcement Learning. *Frontiers in Neuroscience*, *6*(APR). https://doi.org/10.3389/fnins.2012.00043

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Weber, C., Bellebaum, C. (2024) Prediction-error-dependent processing of immediate and delayed positive feedback. *Scientific Reports* 14, 9674. https://doi.org/10.1038/s41598-024-60328-8

Weismüller, B., & Bellebaum, C. (2016). Expectancy affects the feedback-related negativity (FRN) for delayed feedback in probabilistic learning. *Psychophysiology*, *53*(11), 1739–1750. https://doi.org/10.1111/psyp.12738

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic Description of Factorial Models for Analysis of Variance. *Applied Statistics*, *22*(3), 392. https://doi.org/10.2307/2346786

Williams, C. C., Hassall, C. D., Trska, R., Holroyd, C. B., & Krigolson, O. E. (2017). When theory and biology differ: The relationship between reward prediction errors and expectancy. *Biological Psychology*, *129*(June), 265–272. https://doi.org/10.1016/j.biopsycho.2017.09.007

Supplement

8.1. Ratings

Self-report data confirmed that the participants were engaged with the task and paid attention to both the cue and the outcome. Most participants reported to have paid close attention to the cue informing about reward probability (*M* = 5.21, *sd* = 1.44, *Range* 1-7), as well as the stimulus informing about the outcome (*M* = 5.56, *sd* = 1.35, *Range* 1-7). These ratings are comparable with those reported in Hajcak et al. (2003), where participants rated the attention directed to the cue at *M* = 5.69 (*sd* = 1.14), and outcome at *M* = 5.50 (*sd* = 1.32).

8.2. Preprocessing according to current standards & Quantification Methods

To provide another robustness test (3 & 4), we additionally preprocessed the data according to current standards (e.g., using ICA instead of Regression based Ocular correction, using all available electrodes instead of only 5 selected ones, etc.). Moreover, given that Peak Quantifications of ERPs are more sensitive to noise, we also used a (time-window) Mean Amplitude approach (robustness test 2 & 4). Nevertheless, data quality turned out to be comparable across quantifications and measurements. Standardized Measurement Error (SME, according to Luck et al. 2021) are summarized in Supplementary Table 1. As expected, SME values were higher for Peak compared to Mean quantification, and Difference Measures compared to keeping reward and no-reward outcomes separate. The SME values were comparable for the preprocessing following the original study and the one done according to current standards. The ERP waves were also comparable, see Supplementary Table 2, Figures 1 and 3.

Supplementary Table 1: *Mean SME for different Quantification Methods and ERP components*

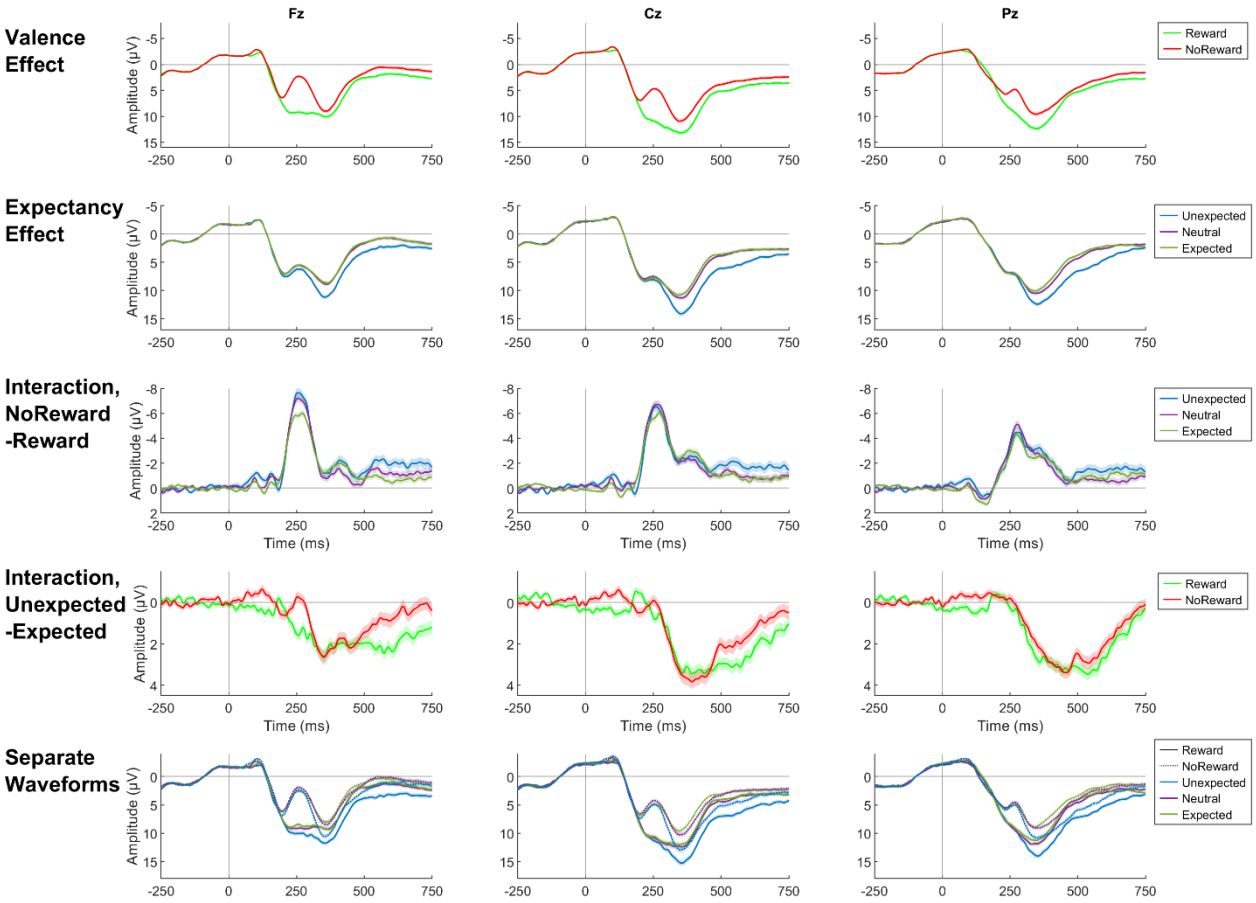
	Peak		Mean	
Preprocessing	Difference	Separate	Difference	Separate
FRN Component				
Original	3.09 (1.45)	2.34 (1.23)	2.59 (1.16)	1.97 (0.94)
Current Std.	3.33 (1.36)	2.47 (1.08)	2.78 (1.15)	2.09 (0.91)
P300 Component				
Original		2.24 (1.16)		2.02 (1.13)
Current Std.		2.29 (0.98)		2.04 (0.90)

Note. Mean SME across Expectancy/Valence/Electrode Levels, with *sd* in parenthesis, for peak and mean quantification of the two ERP components under consideration. Separate/Difference refers to keeping outcome Valence separately (Reward vs. No-Reward) or creating a difference wave (No-Reward minus Reward).

Supplementary Table 2: *Mean SME for different Quantification Methods and ERP components*

		Difference			NoReward			Reward		
		UX	NE	EX	UX	NE	EX	UX	NE	EX
FRN Component										
Original	Peak	-9.65 (6.46)	-9.03 (5.6)	-7.58 (4.57)	1.19 (5.21)	0.71 (5.03)	1.62 (4.63)	8.23 (7.26)	7.24 (5.51)	6.71 (5.79)
Original	Mean	-4.59 (5.16)	-4.91 (4.54)	-4.07 (3.84)	5.33 (5.13)	4.27 (4.8)	4.69 (4.51)	11.04 (6.4)	9.98 (5.7)	9.45 (5.51)
Current Std.	Peak	-10.56 (6.62)	-9.3 (5.48)	-8.14 (4.83)	-0.78 (6.01)	-0.36 (5.28)	0.35 (4.83)	5.32 (6.99)	5.68 (6.01)	5.65 (5.93)
Current Std.	Mean	-4.54 (5.53)	-4.84 (4.55)	-4.17 (3.91)	3.93 (5.33)	3.29 (4.66)	3.77 (4.36)	9.79 (6.48)	8.97 (5.8)	8.66 (5.44)
P300 Component										
Original	Peak				15.42 (7.31)	13 (6.47)	12.51 (6.65)	17.9 (7.66)	15.58 (7.01)	14.5 (6.51)
Original	Mean				10.06 (7.05)	7.97 (6.18)	7.51 (6.02)	12.36 (6.92)	10.03 (6.24)	9.2 (5.94)
Current Std.	Peak				15.03 (6.81)	12.29 (5.69)	11.69 (5.27)	17.38 (7.39)	14.97 (6.52)	13.93 (6.2)
Current Std.	Mean				9.02 (6.17)	7.13 (5.09)	6.49 (4.52)	11.08 (6.4)	8.94 (5.36)	8.35 (5.18)

Note. Mean ERP, with *sd* in parenthesis, for peak and mean quantification of the ERP components by keeping Valence separately (Reward vs. No-Reward) or creating a difference wave (No-Reward minus Reward). UX = Unexpected. NE = Neutral. EX = Expected



Supplementary Figure 1. *ERP Plots using the preprocessing following current standards at electrode sites Fz, Cz, and Pz, separately for the different conditions. Shaded Areas represent +- SEM.*

8.3. Robustness Tests using Bayesian MLMs (Robustness Test 1 – 4)

All data and R scripts can be found on OSF (https://osf.io/xt4c6/). Posterior estimates for the fixed parameters showed convergence, as evidenced by R-hat values below 1.008 across all parameters and all models (different ERPs, Preprocessing, Quantification Methods, see Supplementary Table 3-5). Across all models, the marginal R^2, accounting for fixed effects, ranged from 0.06 to 0.26 indicating that fixed effects alone explained on average 11.97% of the variance in the ERP variations. The conditional R^2, accounting for both fixed and random effects, ranged from 0.89 to 0.98, indicating that the total model explained on average 92.81% of the variance.

Supplementary Table 3: *Estimates of the posterior distributions of the model parameters for the different Robustness Tests and FRN/RewP component*

	Original		Current Standards	
	Peak (RobTest 1)	Mean (RobTest 2)	Peak (RobTest 3)	Mean (RobTest 4)
Intercept	-10.79 (0.43) [-11.61, -9.89] 1.01	-5.86 (0.37) [-6.56, -5.11] 1.01	-11.96 (0.46) [-12.87, -11.07] 1	-5.83 (0.4) [-6.61, -5.05] 1
Location: PZ	2.98 (0.33) [2.34, 3.63] 1	2.74 (0.26) [2.22, 3.23] 1	3.48 (0.32) [2.87, 4.11] 1.01	2.92 (0.28) [2.36, 3.48] 1
Location: CZ	0.82 (0.21) [0.42, 1.22] 1	0.65 (0.15) [0.37, 0.94] 1	1.1 (0.21) [0.68, 1.52] 1	0.94 (0.16) [0.63, 1.25] 1
Expectancy: Neutral	1.26 (0.35) [0.58, 1.93] 1.01	0.11 (0.29) [-0.46, 0.67] 1	1.85 (0.36) [1.13, 2.56] 1	0.13 (0.31) [-0.47, 0.72] 1
Expectancy: Expected	2.64 (0.33) [1.98, 3.26] 1.01	0.99 (0.28) [0.43, 1.54] 1.01	3.18 (0.33) [2.53, 3.82] 1	0.95 (0.29) [0.37, 1.52] 1
Location: PZ Expectancy: Neutral	-0.93 (0.31) [-1.54, -0.31] 1	-0.58 (0.25) [-1.07, -0.09] 1	-0.92 (0.35) [-1.63, -0.26] 1	-0.61 (0.29) [-1.18, -0.05] 1
Location: CZ Expectancy: Neutral	-0.64 (0.21) [-1.06, -0.24] 1	-0.51 (0.15) [-0.8, -0.21] 1	-0.68 (0.22) [-1.11, -0.24] 1	-0.68 (0.17) [-1.01, -0.35] 1
Location: PZ Expectancy: Expected	-0.99 (0.3) [-1.58, -0.4] 1	-0.65 (0.24) [-1.12, -0.18] 1	-1.21 (0.3) [-1.81, -0.61] 1	-0.77 (0.26) [-1.29, -0.26] 1
Location: CZ Expectancy: Expected	-0.95 (0.21) [-1.36, -0.53] 1	-0.67 (0.15) [-0.97, -0.37] 1	-1.05 (0.23) [-1.5, -0.61] 1	-0.92 (0.17) [-1.26, -0.57] 1

Note: First entry corresponds to Mean (standard deviation), second row shows [95 % Confidence intervals] and last entry corresponds to Rhat.

Supplementary Table 4: *Estimates of the posterior distributions of the model parameters for the different Robustness Tests and P300 component*

	Original		Current Standards	
	Peak (RobTest 1)	Mean (RobTest 2)	Peak (RobTest 3)	Mean (RobTest 4)
Intercept	18.22 (0.59)	12.64 (0.54)	17.39 (0.5)	11.08 (0.42)
	[17.02, 19.39]	[11.56, 13.68]	[16.41, 18.37]	[10.27, 11.91]
	1	1	1	1
Valence: NoReward	-2.56 (0.37)	-2.29 (0.4)	-2.34 (0.38)	-2.02 (0.36)
	[-3.28, -1.84]	[-3.07, -1.48]	[-3.09, -1.59]	[-2.73, -1.29]
	1	1	1	1
Expectancy: Neutral	-2.38 (0.27)	-2.4 (0.25)	-2.42 (0.27)	-2.13 (0.27)
	[-2.9, -1.85]	[-2.89, -1.92]	[-2.97, -1.89]	[-2.66, -1.61]
	1	1	1	1
Expectancy: Expected	-3.44 (0.24)	-3.23 (0.22)	-3.47 (0.27)	-2.73 (0.24)
	[-3.92, -2.95]	[-3.66, -2.78]	[-4, -2.94]	[-3.2, -2.25]
	1	1	1	1
Valence: NoReward Expectancy: Neutral	-0.08 (0.34)	0.28 (0.29)	-0.31 (0.38)	0.24 (0.37)
	[-0.75, 0.59]	[-0.29, 0.86]	[-1.06, 0.44]	[-0.48, 0.97]
	1	1	1	1
Valence: NoReward Expectancy: Expected	0.41 (0.34)	0.45 (0.31)	0.13 (0.36)	0.19 (0.34)
	[-0.27, 1.07]	[-0.15, 1.06]	[-0.57, 0.82]	[-0.47, 0.86]
	1	1	1	1

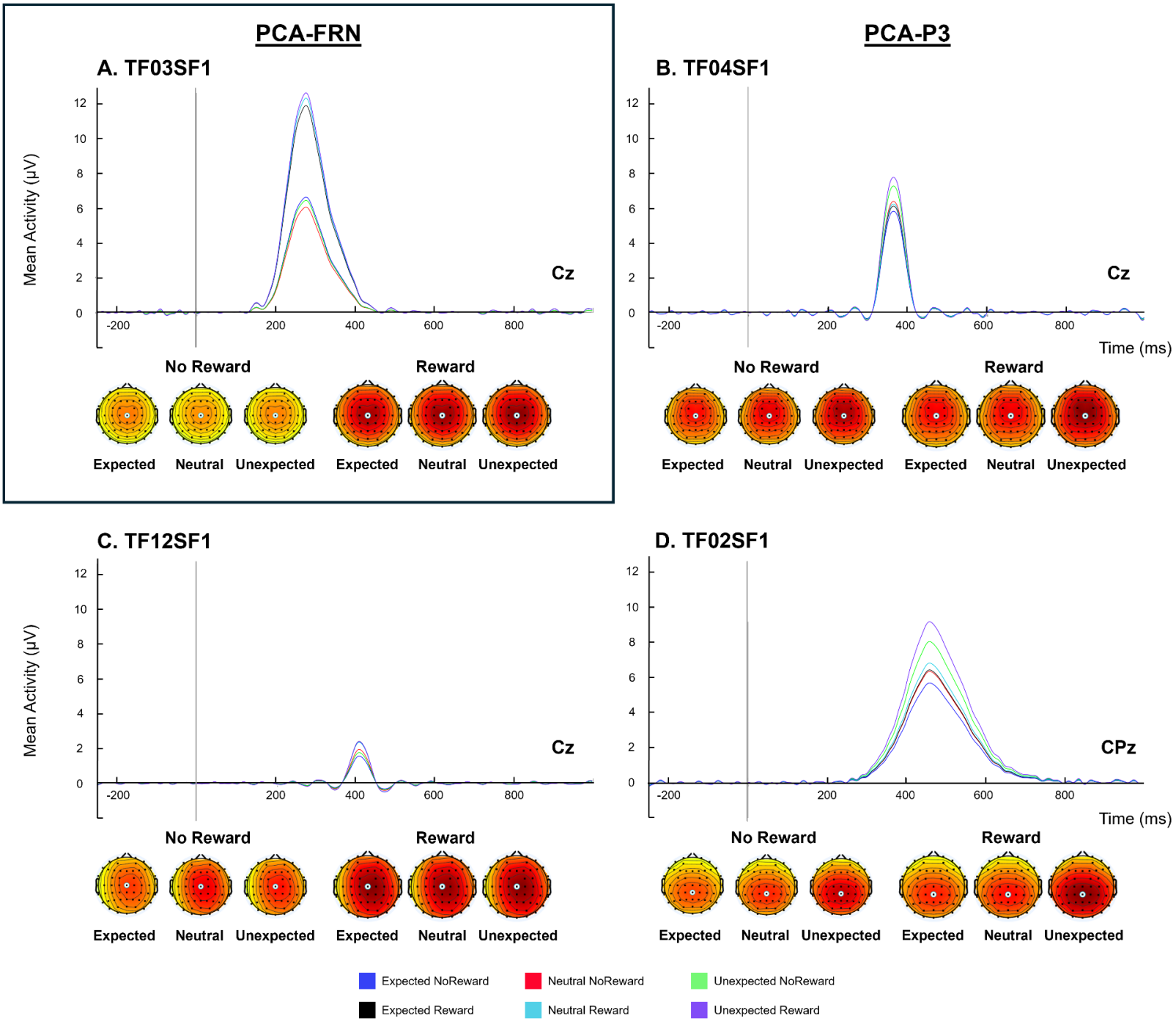
Note: First entry corresponds to Mean (standard deviation), second row shows [95 % Confidence intervals] and last entry corresponds to Rhat.

Supplementary Table 5: *Estimates of the posterior distributions of the model parameters for the different Robustness Tests and FRN component at FZ, keeping Valence separate*

	Original		Current Standards	
	Peak (RobTest 1)	Mean (RobTest 2)	Peak (RobTest 3)	Mean (RobTest 4)
Intercept	8.74 (0.6)	11.27 (0.59)	5.47 (0.56)	9.76 (0.52)
	[7.55, 9.93]	[10.11, 12.45]	[4.35, 6.59]	[8.72, 10.8]
	1	1	1	1
Valence: NoReward	-7.36 (0.42)	-5.84 (0.33)	-6.12 (0.44)	-5.79 (0.39)
	[-8.18, -6.52]	[-6.47, -5.22]	[-6.96, -5.25]	[-6.54, -5.02]
	1	1	1	1
Expectancy: Neutral	-0.97 (0.29)	-1.07 (0.22)	0.37 (0.27)	-0.81 (0.22)
	[-1.55, -0.41]	[-1.51, -0.64]	[-0.17, 0.89]	[-1.23, -0.37]
	1	1	1	1
Expectancy: Expected	-1.3 (0.29)	-1.57 (0.21)	0.33 (0.26)	-1.11 (0.2)
	[-1.85, -0.73]	[-1.99, -1.16]	[-0.19, 0.84]	[-1.51, -0.71]
	1	1	1	1
Valence: NoReward Expectancy: Neutral	0.73 (0.39)	0.09 (0.3)	0.03 (0.36)	0.14 (0.31)
	[-0.03, 1.49]	[-0.48, 0.66]	[-0.69, 0.73]	[-0.46, 0.72]
	1	1	1	1
Valence: NoReward Expectancy: Expected	1.68 (0.37)	0.97 (0.27)	0.88 (0.35)	0.93 (0.27)
	[0.96, 2.41]	[0.45, 1.51]	[0.18, 1.58]	[0.39, 1.46]
	1	1	1	1

Note: First entry corresponds to Mean (standard deviation), second row shows [95 % Confidence intervals] and last entry corresponds to Rhat.

1.4. Robustness through PCA (Robustness Test 6)



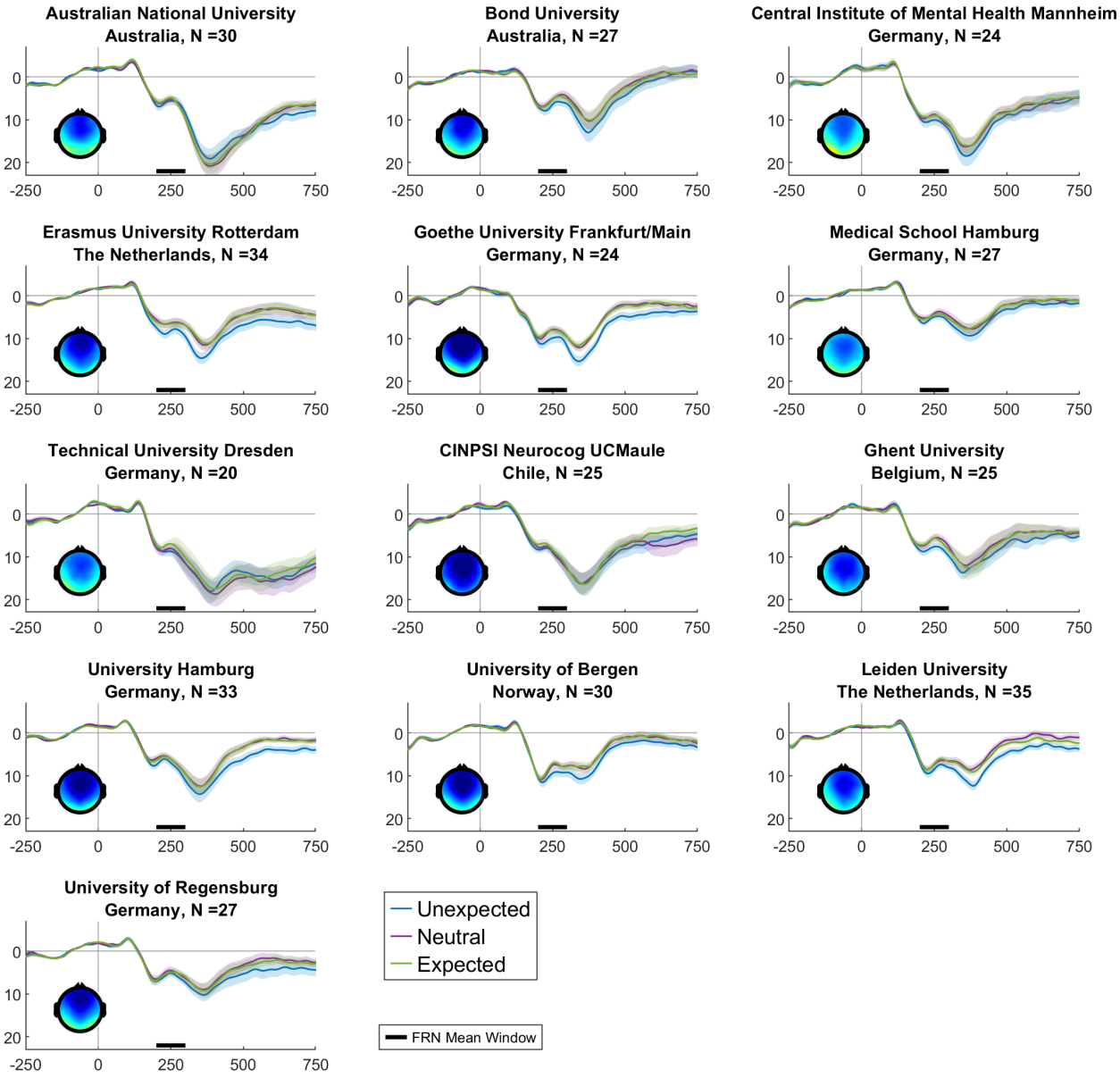
***Supplementary Figure 2**. Activation over time and topographical plots for PCA factors resembling the RewP and P300 components: (A) Factor TF03SF1 (corresponding to the RewP) peaks at 276 ms at the central area. (B) Factor TF04SF1 (corresponding to the P3) peaks at 366 ms at the central area. (C) Factor TF12SF1 (corresponding to the P3) peaks at 410 ms at the central area. (D) Factor TF02SF1 (corresponding to the P3) peaks at 458 ms at the centro-parietal area.*

The PCA factor TF12SF1, corresponding to the P300 component, exhibited a peak latency at 410 ms over the central area (maximal at Cz). The robust ANOVA revealed a significant main effect of Valence ($T_{WJt}/c_{1.0,198.0}$ = 17.48, p < .001, MSe = 6.81), showing a larger positivity for reward than no-reward outcomes (M_{Reward} = 2.34 µV, sd = .01 vs. M_{NoRewad} = 1.71 µV, sd = .02). Both the main effect of Expectancy ($T_{WJt}/c_{2.0,176.0}$ = 1.18, p = .31, MSe = 2.58, and the interaction effect did not reach significance ($T_{WJt}/c($2.0,176.0) = 2.22, p = .111, MSe = 2.57).

The PCA factor TF02SF1, also accounting for the P300 component, exhibited a peak latency at 458 ms over the centro-parietal area (maximal at CPz). The robust ANOVA revealed a significant main effect of Valence ($T_{WJt}/c_{1.0,198.0}$ = 17.11, p < .001, MSe = 10.59) showing a larger positivity for reward than no-reward outcomes (M_{Reward} = 7.33 µV, sd = .02 vs. M_{NoReward} = 6.55 µV, sd = .02). A significant main effect of Expectancy ($T_{WJt}/c_{2.0,176.0}$ = 67.61, p < .001, MSe = 10.80) was also found, explained by a larger positivity for unexpected than expected outcomes ($M_{\text{Unexpected}}$ = 8.47 µV, SD = .03 vs. M_{Expected} = 5.90 µV, SD = .02 vs. M_{Neutral}= 6.45 µV, SD = .02). The interaction between Valence and Expectancy was not significant ($T_{WJt}/c_{2.0,176.0}$ = 1.82, p = .164, MSe = 10.18).

1.5. Lab Effects

Lab Effects were only modeled in the MLMs and the Meta-Analysis, indicating that there was some variation for some effects (e.g., the main effect of Expectancy for the FRN). Although all Labs showed canonical FRN and P300 components and ERP waveforms, there was some variation in the magnitude and timing of these components, see supplementary Figure 3.



***Supplementary Figure 3**. ERP Plots using the preprocessing following the original preprocessing at electrode site Fz, separately for each Lab and Expectancy level (across Valence). Shaded Areas represent +- SEM. Inline of the topographical plots (based on the preprocessing according to current standards), defined as the average amplitude in the 200-300 ms (NoReward - Reward, across expectancy levels).*

Supplementary Table 6: *Mean ERP, Latency and SME values for different Labs and ERP components*

	FRN			P300		
Lab	ERP	Latency	SME	ERP	Latency	SME
ANU	-7.14 (4.21)	269.97 (27.63)	3.19 (1.27)	14.62 (5.37)	376.63 (34.92)	2.13 (0.95)
BON	-8.74 (4.28)	277.34 (36.3)	3.37 (0.51)	14.87 (6.33)	344.93 (49.75)	2.63 (0.55)
CIM	-8.07 (4.51)	272.66 (36.26)	3 (0.83)	17.56 (6.23)	377.28 (51.74)	2.12 (0.85)
ERA	-9.61 (4.66)	271.84 (27.49)	3.34 (1.38)	16.74 (6.59)	364.99 (52.47)	2.53 (1.51)
GUF	-10.02 (4.94)	263.17 (16.69)	3.16 (0.73)	15.23 (4.72)	339.14 (36.57)	2.25 (0.56)
MSH	-7.08 (4.2)	271.29 (28.21)	3.2 (2.45)	11.33 (4.92)	354.17 (43.55)	2.16 (0.99)
TUD	-7.69 (2.91)	274.02 (24.87)	2.5 (0.59)	16.93 (5.56)	369.17 (44.07)	1.58 (0.32)
UCM	-9.48 (6.77)	258.5 (22.21)	3.63 (0.86)	14.61 (8.76)	353.47 (53.37)	2.83 (1.33)
UGE	-8.13 (4.38)	258.85 (33.61)	3.3 (1.37)	15.55 (7.17)	365.89 (57.01)	2.42 (0.89)
UHH	-9.22 (3.91)	262.36 (31.09)	3.24 (0.67)	16.48 (7.18)	346.00 (41.95)	2.47 (0.50)
UIB	-8.29 (3.78)	260.85 (25.15)	2.5 (0.43)	13.07 (4.73)	318.11 (51.05)	1.66 (0.34)
UNL	-8.55 (4.16)	280.52 (27.87)	3.09 (0.58)	13.62 (5.58)	365.54 (41.29)	2.30 (0.43)
URE	-8.5 (3.33)	264.33 (25.44)	2.6 (0.65)	12.96 (5.48)	348.18 (56.65)	1.88 (0.46)

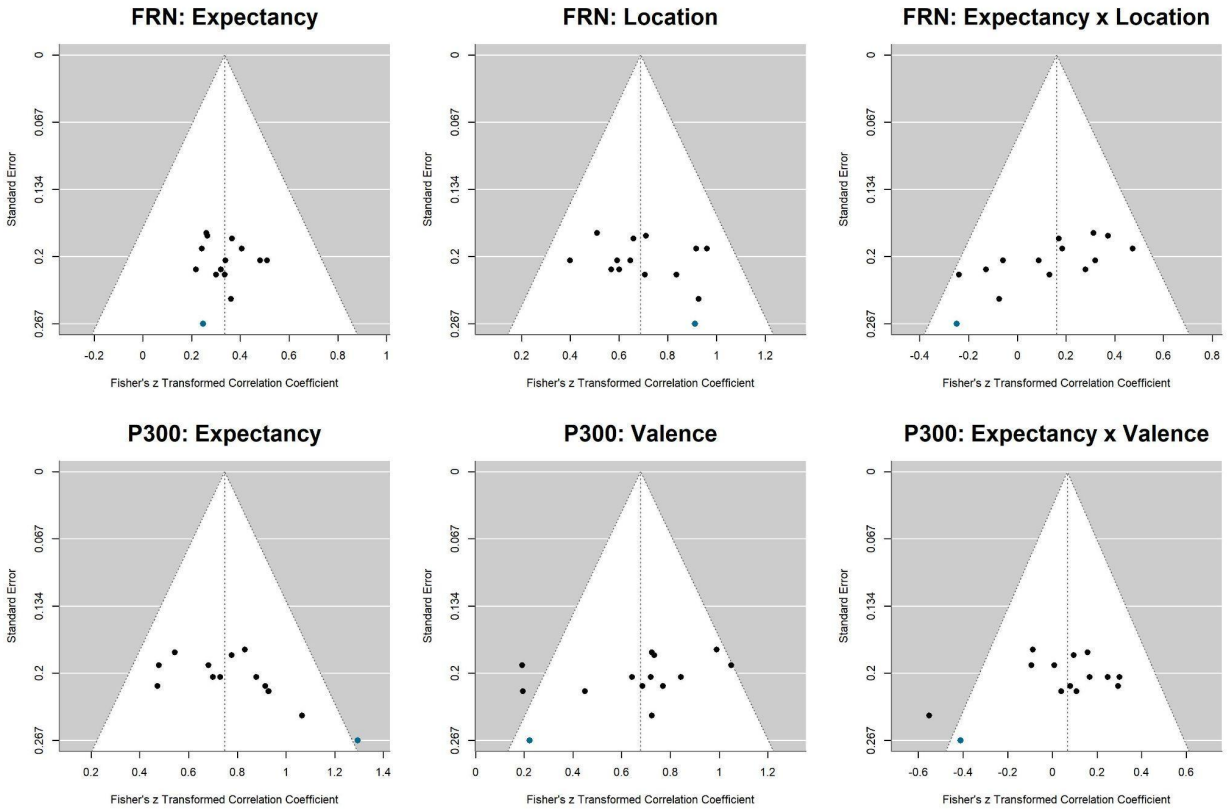
Note. Mean ERP (in µV), Latency (in ms) and SME across Expectancy/Valence/Electrode Levels, with *sd* in parenthesis, for the direct replication (original preprocessing, peak quantification, difference waves for FRN).
ANU = Australian National University, Australia. BON = Bond University, Australia. CIM = Central Institute of Mental Health Mannheim, Germany. ERA = Erasmus University Rotterdam, The Netherlands. GUF = Goethe University Frankfurt am Main, Germany. MSH = Medical School Hamburg, Germany. TUD = Technical University Dresden, Germany. UCM = CINPSI Neurocog UCMaule, Chile. UGE = Ghent University, Belgium. UHH = University Hamburg, Germany. UIB = University of Bergen, Norway. UNL = Leiden University, The Netherlands. URE = University of Regensburg, Germany

Supplementary Table 7. *Overview of EEG set-up and recording details at each replicating lab*

Lab	Amplifier System	Electrode/Cap Model, Number EEG + external electrodes	Sampling Rate	Reference, Ground	acquisition filter bandwidth	operating system (e.g., Windows, Linux, MacOS)	Screen Type, Size, Ratio, Refresh Rate	Stimulus Presentation, Language	Buttons for task	Recording of Resting
Australian National University, Australia	Biosemi	Biosemi, active, 64 + 6	1024	CMS/DRL	LP filter: 5th order CIC at 204 Hz -3dB	Windows 10	LCD, 24 in, 1920:1080, 60 Hz	Presentation (23.1), English	ZCBM on QWERTY keyboard	no
Bond University, Australia	Biosemi	Biosemi, active, 32 + 6	2048	CMS/DRL	LP filter: 5th order CIC -3dB, at 1/5 of sample rate	Windows 10	LCD, 23 in, 1980:1080, 120 Hz	PsychoPy (21.2.3), English	ZCBM on QWERTY keyboard	yes
Central Institute of Mental Health Mannheim, Germany	BrainProducts actiCHamp	BrainProducts actiCap slim/snap, active, 64 + 4	500	Cz, AFz	High and low pass filter 0.1 - 100Hz	Windows 10	LCD, 24 in, 1980:1080, 60 Hz	Psychopy (22.1.3), German	YCBM on QWERTZ keyboard	yes
CINPSI Neurocog UCMaule, Chile	Biosemi	Biosemi, active, 64 + 6	2048	CMS/DRL	LP filter: 5th order CIC at 102 Hz -3dB	Windows 7	LCD, 24 in, 1920:1080, 75 Hz	Psychopy (22.1.3), Spanish	left/right Ctrl/Alt on QWERTZ keyboard	yes
Erasmus University Rotterdam, The Netherlands	Biosemi	Biosemi, active, 64+ 6	512	CMS/DRL	LP filter: 5th order CIC at 102Hz -3dB	Windows 10	LED, 24 in, 1920:1080, 120 Hz	Presentation (23), Dutch	ZCBM on QWERTY keyboard	yes
Ghent University, Belgium	Biosemi	Biosemi, active, 64+6	512	CMS/DRL	LP filter: 5th order CIC at 102Hz -3dB	Windows 10	CRT, 19 in, 1024:768, 75 Hz	Presentation (23), Dutch	ZCBM on QWERTY keyboard	yes

Goethe University Frankfurt am Main and DIPF, Germany	BrainProducts actiCHamp Plus	EasyCap, Custom, actiCap snap, active, 64	500	Cz, FCz	Low cuttoff (s) 10, High cuttoff (Hz) 100	Windows 10	LCD, 24 in, 1920: 1080, 60 Hz	PsychoPy (22.1.3), German	left/right Ctrl/Alt on QWERTZ keyboard	yes
Leiden University, The Netherlands	Biosemi	Biosemi, active, 64+6	1024	CMS/DRL	LP filter: 5th order CIC at 102Hz -3dB	Windows 10	LCD, 24 in, 1680:1050, 60 Hz	Psychopy (22.1.1), Dutch & English	ZCBM on QWERTY keyboard	yes
Medical School Hamburg, Germany	BrainProducts BrainAmp DC	BrainProducts actiCap snap active, 32	1000	FCz, AFz	Low cutoff (s): 10, High cutoff (Hz): 1000	Windows 10	LCD (LED backlight), 23 in, 1920:1080, 60 Hz	Presentation 20.1 (Doors task), Resting-state (Psychopy (22.1.3)), German	YCBM on QWERTZ keyboard	yes
Technical University Dresden, Germany	BrainProducts BrainAmp MR Plus	EasyCap, BrainCap with Multitrodes passive, 64	500	AFF1h, AFF2h	Low cutoff (s): 10, High cutoff (Hz): 1000	Windows 10	LED, 24 in, 1920:1080, 144 Hz	Presentation (19.0), German	YCBM on QWERTZ keyboard	yes
University Hamburg, Germany	Biosemi	Biosemi, active, 64+6	512	CMS/DRL	LP filter: 5th order CIC at 102Hz -3dB	Windows 7	LCD, 24 in, 16:19, 60 Hz	Psychopy (22.1.3), German	left/right Ctrl/Alt on QWERTZ keyboard	yes
University of Bergen, Norway	BrainProducts BrainAmp MR Plus	EasyCap M24 for multitrodes, passive, 32	500	FCz, AFz	Low cutoff (s): 10, High cutoff (Hz): 250	Windows 10	LED, 24 in, 1920:1080,120 Hz	Psychopy (22.1.3), Norwegian	left/right Ctrl/Alt on QWERTY keyboard	no
University of Regensburg, Germany	Bittium NeuroOne Tesla	EasyCap (32 Ch BrainCap for TMS with Multitrodes), passive, 32	1000	FCz, AFz	High cutoff: 250 Hz	Windows 10	LCD, 24 in, 16:19, 60 Hz	Doors task: Psychopy (22.1.3); Resting State: Presentation, German	left/right Ctrl/Alt on QWERTZ keyboard	yes

1.6. Meta Analysis - Funnel Plots



***Supplementary Figure 4**. Funnel plot of the meta-analysis. Each plotted point represents the standard error and standardized Fisher's z Transformed Correlation Coefficient for a single lab. The white triangle represents the region where 95% of the data points would lie in the absence of a publication bias. The vertical line represents the average standardized mean effect. The blue dot represents the effects from the original study.*

1.7. Sample Size Recommendations

Our large sample allowed us to determine precisely the effect sizes for the FRN/RewP and P300 components in a way that could not be achieved in previous EEG studies focused on them, because smaller sample sizes were used. Hence, this information could be extremely valuable as it could guide sample size estimations in future EEG studies on them. However, it is important that we aimed to replicate the ERP effects of Hajcak et al. (2005) using the same experimental procedure, pre-processing and quantification method, and hence adjustments are needed depending on the specific methodology, needs and goals of these future studies.

Supplementary Table 8. *Sample size Recommendation for future use*

FRN
Difference Reward – NoReward
Expectancy (3) x Location (3)

Effect	η_p^2	*N recom.*
Expectancy	0.12 [0.08, 0.17]	**38** [58 – 26]
Location	0.29 [0.23, 0.34]	**14** [18 - 12]
Location x Expectancy	0.01 [≤ 0.01, 0.03]	**298** [n.a. - 98]

FRN at Fz
Valence (2) x Expectancy (3)

Effect	η_p^2	*N recom.*
Valence	0. 66, [0.6, 0.71]	**6** [8 – 6]
Expectancy	0.02 [≤ 0.01, 0.05]	**238** [n.a. - 94]
Valence x Expectancy	0.04 [0.01, 0.07]	**118** [480 - 66]

P300
Valence (2) x Expectancy (3)

Effect	η_p^2	*N recom.*
Valence	0.35 [0.27, 0.42]	**18** [24 – 14]
Expectancy	0.4 [0.34, 0.44]	**10** [12 - 8]
Valence x Expectancy	≤ 0.001 [≤ 0.01, 0.02]	n.a. [n.a. - 238]

Note: η_p^2 from the direct replication (original preprocessing, peak measures) including 95% CI. For a conservative approach, the lower bound of the confidence interval should be used. *N recom* = recommended sample size based on a-priori power analysis in MorePower for α = .05 and β = 80. MorePower returns n.a. if sample size would exceed 2500 participants.