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], 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[20,21], Gilles Pourtois[22]

[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] Department of Psychology, Ural Federal University, Yekaterinburg, Russia
[21] Institute of Medical Psychology and Behavioral Neurobiology, University of Tuebingen, Tuebingen, Germany
[22] Department of Experimental Clinical & Health Psychology, Ghent University, Ghent, Belgium

* Correspondence to Katharina Paul, University of Hamburg, katharina.paul@uni-hamburg.de

Author contributions:

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

Katharina Paul: Data curation, Formal Analysis, Funding acquisition, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing

Douglas J. Angus: Supervision, Writing – review & editing

Florian Bublatzky: Funding acquisition, Supervision, Writing – review & editing

Raoul Dietrich: Methodology, Investigation, Writing – review & editing

Tanja Endrass: Supervision, Writing – review & editing

Lisa-Marie Greenwood: Investigation, Supervision, Writing – review & editing

Greg Hajcak: Writing – review & editing

Bradley N. Jack: Investigation, Supervision, Writing – review & editing

Sebastian P. Korinth: Investigation, Writing – review & editing

Leon O. H. Kroczek: Investigation, Supervision, Writing – review & editing

Boris Lucero: Funding acquisition, Investigation, Writing – review & editing

Annakarina Mundorf: Investigation, Writing – review & editing

Sophie Nolden: Supervision, Writing – review & editing

Jutta Peterburs: Funding acquisition, Methodology, Supervision, Writing – review & editing

Daniela M. Pfabigan: Funding acquisition, Supervision, Validation, Writing – review & editing

Antonio Schettino: Formal Analysis, Visualization, Writing – review & editing

Yee Lee Shing: Funding acquisition, Supervision, Writing – review & editing

Gözem Turan: Investigation, Writing – review & editing

Melle J.W. van der Molen: Supervision, Writing – review & editing

Matthias J. Wieser: Investigation, Supervision, Writing – review & editing

Niclas Willscheid: Investigation

Faisal Mushtaq: Conceptualization, Project administration, Writing – review & editing

Yuri G. Pavlov: Conceptualization, Funding acquisition, Project administration, Writing – review & editing

Gilles Pourtois: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing

Abstract

Two event-related brain potential (ERP) components elicited during feedback processing are the frontocentral feedback-related negativity (FRN), followed by the posterior P300. According to the Error-Related Reinforcement Learning Theory (Holroyd & Coles, 2002), the FRN amplitude is largest when the outcome is negative and unexpected. Complementing this, studies on the subsequent P300 have often reported larger amplitudes for positive than negative outcomes. In an influential ERP study, Hajcak et al., (2005) manipulated outcome valence and expectancy in a guessing task. However, they found that the FRN component was larger for negative (no-reward) than positive (reward) outcomes, irrespective of expectancy. Conversely, the P300 component was larger for unexpected than expected outcomes, irrespective of valence. These results were at odds with prominent theories and extant literature. Here, we aim to replicate these results within the #EEGManyLabs project (Pavlov et al., 2021). Across thirteen labs we will not only undertake a close replication, but test the robustness of these effects to analytical choices (e.g. quantification of ERPs) and supplement the findings with Bayesian multilevel linear models to test for the reported absence of the effects.

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 have 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 Error-Related Negativity Reinforcement Learning Theory (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). Importantly, this theoretical model assumes that the FRN is larger for worse-than-expected relative to better-than-expected outcomes, i.e., the FRN is largest for negative prediction errors (see also Walentowska et al., 2019; Walsh & Anderson, 2012). Moreover, whether the feedback is utilitarian (e.g., incentive-related) or performance-related (e.g., informing about accuracy) is irrelevant, as this signed prediction error captured by the FRN is equally large for unexpected negative outcome 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 not expected (50%) outcomes, and for not expected 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 remain hotly 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 influential 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 in part be the product of a superimposed positive-going deflection, a so-called Reward Positivity (RewP; see Proudfit, 2015). Accumulating evidence shows that the RewP could capture different performance monitoring or motivational effects than the FRN (i.e., consummatory reward processing for the RewP as opposed to reward prediction errors for the FRN; (see Foti et al., 2011; Gable et al., 2021), even though their time-courses and scalp distributions partly overlap (Gheza et al., 2018; Krigolson, 2018). It remains to be seen if the proposed sensitivity of the FRN/RewP is driven by worse-than-expected signals (negative prediction errors) or by better-than-expected signals

(positive prediction errors). 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, next to the FRN and P300 components, it appears important to consider the RewP when studying performance monitoring.

The results of this study sparked numerous conceptual replications on the nature of the FRN and the P300 component across different tasks, motivational contexts, and in clinical and non-clinical populations. To date, the work has been cited over 530 times (Google Scholar in August 2022). Yet, despite this intense focus, there has been no direct replication of the original procedure, measures, and analyses. The goal of the present study is to undertake a multi-lab replication of Hajcak et al. (2005), using a trial-by-trial manipulation of both expectancy and valence. We intend 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 hypothesize that:

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

Finally, if, in contrast to the original replication, but in line with the RL-Theory, we find an effect of expectedness on FRN/RewP amplitudes, we will explore which component is driving this effect.

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 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] are only a rough overestimation of the true effect size. Additionally, there is 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 (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 upper bound of the FRN effect of expectancy during guessing tasks, we refrain 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

[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 $F = 1$.

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 will allow us to identify a much smaller effect than any individual study on this matter has been able to identify so far.

A similar rationale can be 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 will have 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 will be recruited via local advertisements or online recruitment systems. For their participation, they will be reimbursed with 15 EUR/ NOK 300 or course credits. Additional to that, each participant will receive a payout of their in-task wins of 5 EUR/ 17 AUD/ 50 NOK / 5000 CLP. Participants are told that they could increase their payouts if they choose the “correct door”. However, regardless of their choices the outcome is 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 (Ghent University [2022/14]; German Psychological Society (DGPS) [PK-22-02-21]; Bond University [DA03365]; University of Oslo, Department of Psychology [20317283] & NSD [320122]; Leiden University [2022-05-12-M.J.W. van der Molen-V2-3819]; others are in progress).

2.2. Procedure

The procedure will follow the process employed in Experiment 1 in Hajcak et al. (2005) as closely as possible, and any departures from this will be explicitly stated. Participants will be tested individually in an EEG laboratory. Upon their arrival in the lab, they will receive a brief description of the experiment and will provide informed consent. Then they will be prepared for EEG recording and the EEG electrodes will be attached. Participants will be familiarized with the guessing task and the feedback using a practice block consisting of 40 trials (not included in the analysis). Afterwards, they will complete 6 blocks of the guessing task, with each block comprising 40 trials (240 trials in total). Self-paced breaks will be allowed in between blocks. Every other block, the experimenter will enter the testing room to inform about the current winnings (which are presented on the screen), monitor the EEG signal, and keep participants alert.

As this project is part of a wider initiative on replicability in EEG (#EEGManyLabs), most of the laboratories in this replication will also collect resting state data EEG data together with some personality measures (https://osf.io/sp3ck/, (Pavlov et al., 2021). Neither EEG nor personality data will be 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 will record 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 will ask 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 (Events, 1977), and the Short Version of the Big Five Inventory (Gerlitz & Schupp, 2005) questionnaires. In the labs that do not record this additional data (see Supplementary Table 1), only the guessing task is presented.[2]

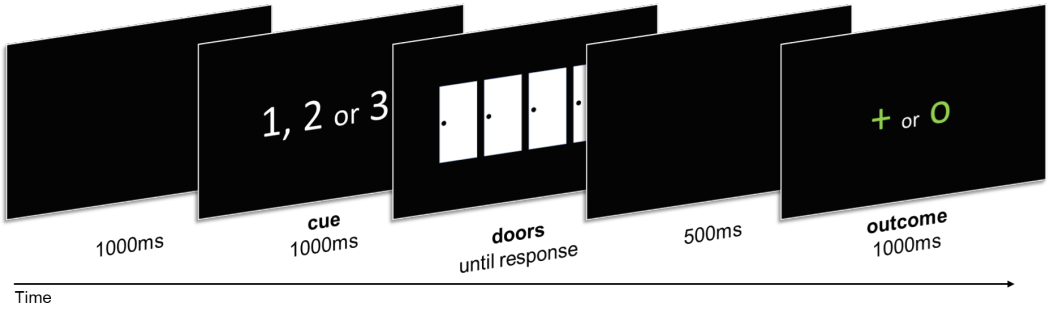

Figure 1: Trial structure. Each trial comprises three successive visual events: a cue (that informs about reward probability in the current trial), followed by the presentation of four doors (imperative stimulus; the participant is asked to pick one of them based on guessing), before the outcome (either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

Each trial starts with a cue presented for 1000 ms in the center of the screen (see Figure 1). The cue is presented as the number 1, 2, or 3, corresponding to a probability of winning of 25%, 50%, or 75%. After this cue, four doors appear in the center of the screen and the participant is asked to select one of them by pressing one of four pre-defined keys on the keyboard (exact keys vary 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 1). Participants are asked to guess which door could contain a prize. The four doors stay on screen until the response/choice. Then a blank screen ensues (500 ms), before the outcome is presented in green font for 1000 ms. The outcome is presented as a “+”, indicating that a small monetary reward is attained (value is 0.04 EUR or 0.15 AUD or 0.4 NKR or 35 CLP), or as a “o”, indicating that no-reward is attained. The trial ends with a 1000 ms blank screen used as inter-trial interval. Stimuli are presented in white on black background. Accordingly, in this task, reward motivation is promoted while no punishment motivation is involved.

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

There are six experimental conditions, corresponding to the combinations of cue and outcome: expected reward (i.e., “+” symbol following “3” used as cue, 60 trials), not expected 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), not expected 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 are shown in random order.

Upon completion of the task, participants will be 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 are answered on a seven-point scale, ranging from “ignored it” to “paid close attention” by the corresponding numbers on the keyboard.

The whole experiment will last approximately 1 to 1.5 hours. The experiment is 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 1.

Supplementary Table 1. *Overview of EEG set-up and recording details at each replicating lab*

Participating University	Amplifier System	Electrode/Cap Model, Number EEG + external electrodes	Sampling Rate	Reference, Ground	Online Filter	Operating System	Screen Type, Size, Ratio, Refresh Rate	Stimulus Presentation, Language	Buttons for task	Recording of additional data (Resting/ Questionnaires)
Australian National University, Australia	Biosemi	Biosemi, active, 64 + 6	512	CMS/DRL	LP filter: 5th order CIC at 102 Hz -3dB	Windows 10	LCD, 24 in, 1920:1080, 60 Hz	Psychopy (20.1.3), English	ZCBM on QWERTY keyboard	yes
Bond University, Australia	Biosemi	Biosemi, active, 32 + 6	512	CMS/DRL	LP filter: 5th order CIC at 102Hz -3dB	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	BrainProducts actiCap slim, active, 64	500	FCz, AFz	High and low pass filter 0.1 - 100Hz	Windows 10	LCD, 24 in, 1980:1080, 60 Hz	Presentation (20.3), German	YCBM on QWERTZ keyboard	yes
CINPSI Neurocog UCMaule, Chile	Biosemi	Biosemi, active, 64 + 6	512	CMS/DRL	LP filter: 5th order CIC at 102 Hz -3dB	Windows 7	LCD, 24 in, 1920:1080, 75 Hz	PsychoPy (20.1.3), 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 (23.0), Dutch	ZCBM on QWERTY keyboard	yes
Goethe University Frankfurt am Main and DIPF, Germany	BrainProducts	EasyCap, active, 64 + 1	1000	FCz, AFz	LP filter: 5th order Butterworth at 250Hz 30dB	Windows 10	LCD, 24 in, 1920: 108, 60 Hz	PsychoPy (21.1.4), German	left/right Ctrl/Alt on QWERTZ keyboard	yes
Leiden University, The Netherlands	Biosemi	Biosemi, active, 64+6	512	CMS/DRL	LP filter: 5th order CIC at 102Hz -3dB	Windows 10	LCD, 24 in, 16:19, 60 Hz	Psychopy (22.1.1), Dutch	ZCBM on QWERTY keyboard	yes
Medical School Hamburg, Germany	BrainProducts	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), German	YCBM on QWERTZ keyboard	yes
Technical University Dresden, Germany	BrainProducts	EasyCap, passive, custom equidistant montage (No. 10), 63 + 1	500	AFF1h, AFF2h	Low cutoff (s): 10 High cutoff (Hz): 250	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 (20.1.3), German	left/right Ctrl/Alt on QWERTZ keyboard	yes

University of Oslo, Norway	BrainProducts	EasyCap No. 3, active, 32 + 1	500	FCz, AFz	LP filter: 5th order Butterworth at 250Hz 30dB	Windows 10	LED, 24 in, 1920:1080,120 Hz	Psychopy (20.1.3), Norwegian	left/right Ctrl/Alt on QWERTY keyboard	no
University of Regensburg, Germany	NeuroOne	EasyCap, passive, 32	1000	FCz,AFz	LP filter: 250 Hz	Windows 10	LCD, 24 in, 16:19, 60 Hz	Psychopy (20.1.3), German	left/right Ctrl/Alt on QWERTZ keyboard	yes

2.3. Neurophysiological recordings

The replicating labs will be 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 will record 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 will be used to record electro-oculogram (EOG), and two on the left and right mastoids. One EOG electrodes will be attached at least below the left eye, additional electrodes might be placed above the left eye and on the outer canthi of the two eyes. The EEG (and EOG) data will be sampled at 512, 500, 1000 Hz (depending on the setup). Labs also vary 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 2.[3]

2.4. Artifact removal and EEG preprocessing

Data preprocessing will closely follow the original study, including the following steps: activity recorded from Fz, Cz, and Pz and the additional external electrodes will be: (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 by Mittner 2007, see attached scripts); (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 follows the one provided in the original study, the data will also be preprocessed according to recent developments in psychophysiology, which will allow us to test the robustness of the results. Activity recorded from all EEG sensors will be: (i) down-sampled to 500/512 Hz (if recorded with higher sampling rates);

[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).

[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 record with a different setup than the original study, we chose this cut-off instead.

(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 50 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 are 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 will be included.

Nevertheless, to test the robustness of the results, we will aim to ensure good data quality in two ways: First, from all complete recordings, we will exclude participants who have more than 75% of trials rejected (i.e., only 60 trials out of the 240 trials used). Second, we will exclude participants who have less than 8 trials per condition (as the FRN shows good internal consistency with at least 8 trials (Ethridge & Weinberg, 2018). Included trial number as well as standardized measurement error (Luck et al., 2021) will be 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 will be 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 will be quantified at Fz, Cz, and Pz as follows: First, a difference wave will be created by subtracting the ERP observed for reward outcomes from the ERP observed for no-reward outcomes. This difference wave will be computed separately for expected outcomes (expected no-reward minus expected reward), not expected outcomes (not expected no-reward minus not expected reward), and unexpected outcomes (unexpected no-reward minus unexpected reward). For each level of expectancy, the FRN will be defined as the maximum negative amplitude of these difference waves within a window between 200 and 500 ms following outcome onset. The P300 will be scored at Pz as follows. Unlike the FRN, no difference wave will be created. For each of the six conditions, the P300 will be 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 will also score the FRN (or alternatively the 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 will allow us to test the robustness of the results. The FRN/RewP will be 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 will be scored as the mean amplitude 300 - 500 ms following outcome onset.

Considering that the FRN and the P300 components occur in rapid succession, we will additionally quantify 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) will be used for this analysis. Considering the differences in the recording systems that will be used, the individual ERPs will first be standardized. Specifically, data will be downsampled to a common denominator (500 Hz) and only common electrodes will be used. The ERPs will be 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 will be used. The procedure will begin 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 will depend on the scree plot such that only factors explaining more variance than identified in random data will be included (similar to parallel testing, see Dien, 2012). From all temporospatial factor combinations, default windowing will be applied to screen out factors explaining less than 0.5% variance. All remaining factors will be reconstructed into voltage space, in which the voltage accounted for at the peak time point and channel are evaluated as ERP waveforms. Factors whose peak latencies and channels will coincide (based on visual inspection) with the canonical scalp distribution and time course of the FRN (fronto-central, 200-300 ms) and P3 components (posterior-central, 300-500 ms) will be tested.

2.7. Statistical Analyses

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

2.8. Statistical Analyses

2.8.1. Direct Replication through ANOVAs

The ERP amplitudes calculated from the preprocessing and quantification methods used in the original study are subjected to two ANOVAs. For the FRN, the peak amplitude values will be analyzed using a 3 (Location) x 3 (Expectancy) ANOVA. For the P300, a 2 (Valence) x 3 (Expectancy) ANOVA will be used. In case a sphericity violation is detected, Greenhouse–Geisser correction will be applied to p values. The significance alpha level will be set to 0.02. If the ANOVA reveals an effect of expectancy for the FRN, we will follow up on this with a 2 (Valence) x 3 (Expectancy) ANOVA on the amplitudes extracted at Fz (where it was shown to be maximal in

the original study). A significant interaction of Valence and Expectancy, and the corresponding post-hoc tests will be used to test if this was driven by the response to reward outcomes (in the sense of a RewP) or no-reward outcomes (in the sense of a FRN).

Table 1. *Overview of Planned Analyses*

Analytical Step	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	Current standard	Current standard	Original	Current Standard
Outlier handling	None	None	None	Applied	Applied	None	Applied
Quantification of ERPs	Peak	Peak	Mean	Peak	Mean	Peak	PCA
Statistical Test	ANOVA	MLM	MLM	MLM	MLM	Meta-Analysis on ANOVA	ANOVA

2.8.2. Robustness test through MLMs

To better account for variability across participants and laboratories, we will fit eight Bayesian multilevel linear models on the FRN and P300 amplitude values. These models are set up identically, but the dependent variable will be 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 crossing these analytical choices, we will be able to assess the impact of these choices on the outcome and the robustness of the replication.

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

FRN_amplitudes = 1 + location * expectancy + (1 + location * expectancy | laboratory / participant)

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

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

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

Robustness test 3. Amplitudes will be extracted after the modernized preprocessing and defined as the maximum peak in the specified time window.

Robustness test 4. Amplitudes will be extracted after the modernized preprocessing and defined as the mean in the specified time window.

We will allow 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 will choose 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 will place a weakly informative prior on intercepts and slopes: a normal distribution with $\mu = 0$ and and $\sigma = 10$. Since we have no prior knowledge regarding the other model parameters (e.g., standard deviation of laboratory or participant), we will keep the software default weakly informative priors.

Models will be fitted in *R* using the *brms* package (Bürkner, 2018), which employs 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 will start sampling by using 4 MCMC chains with 4000 iterations (2000 warm-up) and no thinning. In case of non-convergence, we will increase the number of iterations by 500 until convergence will be reached or a maximum of 8000 iterations per chain. Model convergence will be 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) between 1 and 1.05 (see also Nalborczyk et al., 2019). Goodness-of-fit will be assessed via Bayesian R^2 (Gelman et al., 2019).

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

The existence of an effect will be 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, we will consider 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 will additionally compute Bayes factors (*BF*; (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). Bayes factors will be 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 will descriptively qualify BF 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., $BF_{01} = 1/BF_{10}$) will indicate the corresponding evidence in favor of H_0.

2.8.3. Meta-Analysis (Robustness Test 5)

Even though each replicating lab only has the statistical power to test the effect of expectancy on the P300, the data of each lab separately will be subjected to the same ANOVAs described in (i). Then, a random effects meta-analysis will be run where the effect sizes of valence (for the P300) or electrode (for the FRN), expectancy, and their interaction gathered in each replicating lab will be combined. Forest and funnel plots will be computed. We will report and plot median and distribution of the weighted and unweighted effect sizes, 95% confidence intervals, and the number of labs successfully replicating the original effect. The metafor package (Viechtbauer, 2010) for R will be used for the meta-analysis.

2.8.4. Temporospatial Principal Component Analysis (Robustness Test 6)

The PCA factors will be analyzed using the statistics function of the EP toolkit using all default parameters. The implemented ANOVAs are robust against violations of statistical assumptions. It includes 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 estimates the population distribution instead of assuming the normality of this distribution; and (iii) a Welch–James approximate degrees-of-freedom statistic that does not assume the homogeneity of error variance (Dien, 2010b). The robust 2x3 repeated-measures ANOVA will include the within-subject factors Valence and Expectancy. The p-value will be adjusted with the Bonferroni correction for multiple comparisons. Follow-up tests for significant interactions will be reported. In case the interaction effect needs a better characterization of its source, a robust t-test will be performed in R Studio using the Yuen test (Yuen, 1974) of the WRS2 (Mair & Wilcox, 2020) package. This particular test allows for mean trimming, making the analysis consistent with the parameters implemented in the robust ANOVA of the EP Toolkit.

Similar to the results from the main analyses above, we expect for the factor corresponding to the FRN 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 expect a significant effect of expectancy (more factor positivity for unexpected outcomes), but no effect of valence or their interaction.

2.9 Evaluation of the Replication and Robustness of Effects

The replication's success will mainly be evaluated in the light of the outcomes of the ANOVAs (see (i) above): The FRN results will be considered to be replicated successfully if the ANOVA shows a significant main effect of position (Fz > Pz), but no significant effect of expectancy or the interaction of expectancy and position. The P300 results will be considered to be replicated successfully if the ANOVA shows 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 provide preliminary robustness tests by comparing these results to the outcomes of the MLMs (see (ii) above) and a PCA (see (iv) above). If the MLMs and the PCA provide evidence for a similar pattern of results as (i), the effect will be considered not only to be replicated but robust and, to some extent, independent of analytical choices. If the direct replication fails, i.e. significant effects are detected where none were expected, or expected effects do not reach significance, the MLMs will be particularly important to conclude if the effects are present or not. If the pattern diverges across the robustness tests, possible sources of these discrepancies will be discussed (with regard to preprocessing choices and/or quantification of the ERPs). Finally, the results of the MLM, (Robustness Test 1) will be compared to the meta-analysis (see (iii) above).

2.10. Analysis of ratings

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

2.11. Sharing of Data and Code

Pre-processing steps will be carried out using EEGLAB 2022.0 (Delorme & Makeig, 2004) implemented in MATLAB 2019, while statistical analyses will be carried out in R (R-Core-Team, 2019). All experimental procedures, pre-processing scripts, analytical analyses and the results of the meta-/mega-analysis were tested on pilot data and will be shared openly, using the Open Science Framework (OSF, https://osf.io/2w9gy, ReadOnlyLink for review: https://osf.io/2w9gy/?view_only=d79c0538c9e04f1298848dcfd7266d5d). All collected data will be made available online through GIN (https://gin.g-node.org/).

3. 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.

4. Acknowledgements

#EEGManyLabs is funded by the DFG (PA 4005/1-1), provided to YGP. FB is funded by DFG (BU 3255/1-2). AS is employed at Erasmus Research Services (Erasmus University Rotterdam) as Senior Advisor Open Science. KP is funded by DFG (PA 4014/2-2). DMP is funded by the South-Eastern Norway Regional Health Authority (2021046). JP is funded by DFG (PE 2077/6-1; PE 2077/7-1). YLS is funded by the European Union (ERC-2018-StG-PIVOTAL-758898).

5. 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

Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature*, *508*(7496), 312–313.

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.

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

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

Events, L. (1977). The CES-D Scale : A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, *1*(3), 385–401.

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. ., & 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. (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

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

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

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

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

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

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. (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

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., 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

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

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.

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

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

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

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

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., 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

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

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*(1), 165–170. https://doi.org/10.1093/biomet/61.1.165