# Modulators of the Processing of Own and Observed Actions

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
**Christine Albrecht**
aus Düsseldorf

Düsseldorf, September 2022

aus dem Institut für experimentelle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Christian Bellebaum
2. Prof. Dr. Gerhard Jocham

Tag der mündlichen Prüfung: 25.11.2022

## Content

## Summary

The posterior medial frontal cortex (mPFC) is a major contributor to action monitoring. In response-locked electroencephalography (EEG), two frontocentral event-related components (ERPs) with presumed origins in medial prefrontal cortex (mPFC) are associated with action monitoring: the error-related negativity (ERN) peaking 80-100 ms after error commission, and an observer mediofrontal negativity (oMN) peaking 100-300 ms after error observation. The predicted outcome-response (PRO) model and empirical findings suggest that the mPFC represents not action valence but expectancy violations. Observed action monitoring processes are probably further modulated by empathy, and possibly in relation with expectancies. Effects of subjective and objective error significance on action monitoring ERP components challenge the PRO model as they cannot be explained by expectancies. This PhD project aimed to investigate effects of expectancies, empathy and error severity as potential modulators of observed action processing. Studies 1 and 2 used a false-belief paradigm to differentiate valence and expectancy effects on observed action monitoring and its modulation by empathy. Study 1 showed that an early observed action monitoring component resembling the oMN represents expectancies, not vicarious error processing, and the results suggested an indirect effect of empathy on ERP amplitudes. This was confirmed in study 2 by showing that single-trial expectancy values influenced ERP amplitudes of a later ERP component, and were themselves influenced by empathy. Empathy could not explain additional variance of the ERP component. In study 3, we found (contrary to the PRO model) that error severity modulated the ERN in a piano playing paradigm. However, we found no error severity, but a binary valence effect for the oMN, which might be explained by expectancies. Expanding the PRO model, we suspect that the mPFC sends a need-to-adapt signal if either actions or predictions need to be adapted. Depending on the task, the focus might be either on predictions (studies 1 and 2) or actions (active paradigm in study 3). Future studies might investigate this theory by manipulating action and prediction deviations and the task focus.

## List of Abbreviations

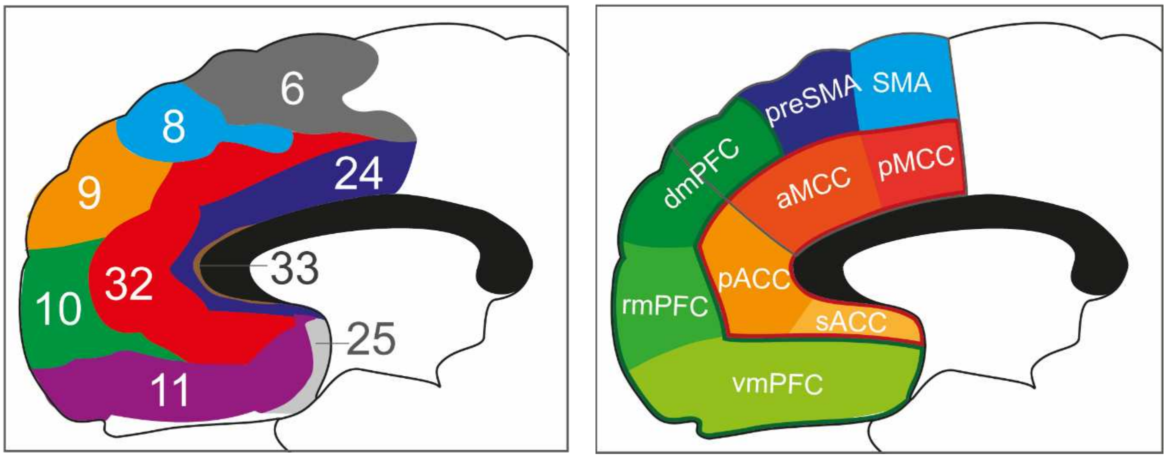
| | |
|---|---|
| ACC | anterior cingulate cortex |
| aMCC | anterior midcingulate cortex |
| BA | Brodmann area |
| BOLD | blood-oxygenation-level dependent |
| dmPFC | dorsomedial prefrontal cortex |
| EEG | electroencephalography |
| ERN/Ne | error-related negativity/error negativity |
| ERP | event-related potential |
| fMRI | functional magnetic resonance imaging |
| FRN | feedback-related negativity |
| IKI | inter-keypress interval |
| IRI | interpersonal reactivity index |
| LME | linear mixed effect |
| MCC | midcingulate cortex |
| MIDI | musical digital interface |
| mPFC | medial prefrontal cortex |
| oFRN | observer feedback-related negativity |
| oMN | observer mediofrontal negativity |
| pMCC | posterior midcingulate cortex |
| pMFC | posterior medial frontal cortex |
| preSMA | pre-supplementary motor area |
| PRO model | predicted response-outcome |
| rewP | reward positivity |
| rmPFC | rostromedial prefrontal cortex |
| SMA | supplementary motor area |
| vmPFC | ventromedial prefrontal cortex |

## Preface: Anatomy of the prefrontal medial wall

Because various terms have been used in previous literature to refer to different parts of the prefrontal medial wall and some definitions remain unclear, I will shortly introduce the anatomy and the terminology that this dissertation is based on. The term medial prefrontal cortex (mPFC) has been used to describe just the isocortical regions of the prefrontal cortex (Etkin et al., 2011), including dorsomedial prefrontal cortex (dmPFC), rostromedial prefrontal cortex (rmPFC), and ventromedial prefrontal cortex (vmPFC; Ullsperger et al., 2014), but it can also refer to a broader area that includes the anterior cingulate cortex (ACC; e.g. Alexander & Brown, 2011). The anterior cingulate cortex, respectively, might refer to the original definition by Brodmann in which the Brodmann Areas (BAs) 24, 25, 32 and 33 were included, or to a smaller area including only the rostral/ventral part as ACC (see Stevens et al., 2011), while the dorsal part of the ACC is referred to as midcingulate cortex (MCC). Ullsperger et al. (2014) refer to the posterior medial frontal cortex (pMFC) as including the rostral and caudal cingulate zone (dorsal and caudal ACC or MCC), pre-supplementary motor area (preSMA) and parts of the dorsomedial prefrontal cortex. Because the subareas of the prefrontal cortex are highly interconnected with each other (see Ridderinkhof, van den Wildenberg, et al., 2004, for a review), it is especially difficult to distinguish between distinct functional areas. To use concise terminology throughout this work, I will refer to the whole area around the prefrontal medial wall as mPFC (including parts of BA 6, 8, 9, 10, 11, 12, and BA 24, 25, 32 and 33). I will further refer to the cingulate areas as ACC (BA 24, 25, 32 and 33) and the isocortical areas as isocortical mPFC (parts of BA 8-12) or directly refer to the subdivisions of the isocortical mPFC. I will use the Brodmann definition of the ACC, but use the terms (from rostral-dorsal to caudal-ventral) posterior MCC (pMCC), anterior MCC (aMCC), pregenual ACC and subcallousal ACC to refer to its subdivisions (according to Ullsperger et al., 2014). The term pMFC thus refers to the more posterior part of the mPFC. Please see Figure 1 for a schematic representation of the respective areas.

**Figure 1**

*Schematic representation of the prefrontal medial wall*




*Note.* On the left: Brodmann areas at the prefrontal medial wall (Brodmann, 1909). On the right: distribution of the prefrontal medial wall (according to Ullsperger et al., 2014). I will refer to all colored areas as the medial prefrontal cortex (mPFC). I have marked what I will call the anterior cingulate cortex (ACC) with a red border, and what I will call isocortical regions of the mPFC with a green border. The ACC consists of the posterior midcingulate cortex (pMCC), anterior midcingulate cortex (aMCC), pregenual ACC (pACC) and subcallousal ACC (sACC). The isocortical regions of the mPFC consist of the dorsomedial prefrontal cortex (dmPFC), rostromedial prefrontal cortex (rmPFC) and ventromedial prefrontal cortex (vmPFC). Above the aMCC and pMCC are the supplementary motor area (SMA) and pre-SMA. The posterior medial frontal cortex (pMFC) contains the aMCC and preSMA and additionally parts of the SMA, pMCC and dmPFC (according to Ullsperger et al., 2014; marked with a grey border).

## Introduction

The human brain likes to be in control, especially when losing control has severe consequences. For example, picture an acrobat being tossed three meters high into the air to do a flip. At this moment, the brains of the acrobat and catchers down on the ground constantly monitor all available information to maximize control: not only do they monitor their own actions, but they monitor others' actions and especially consequential errors. In short, their brains process as much relevant information as possible to maximize the likelihood of a safe landing.

Action monitoring in any form relies heavily on the mPFC, especially the ACC (see Ridderinkhof, Ullsperger, et al., 2004). The mPFC codes negative outcomes (Becker et al., 2014; Holroyd et al., 2004; Nieuwenhuis et al., 2004) as well as own errors (Gawlowska et al., 2018; Holroyd et al., 2004; Ullsperger & Cramon, 2004) and errors observed in others (de Bruijn et al., 2009; Shane et al., 2008; Yoshida et al., 2012). mPFC activity is also linked to subsequent behavior adaptation (Garavan et al., 2002; O'Doherty et al., 2003). While the mPFC is established as control center of the brain, the exact mechanisms of how this control is executed are still unclear, leading to several theories concerning their function.

Specifically, the mPFC has been linked to error monitoring. In the early 1990s, scientists found that a negative-going event-related potential (ERP) component in the electroencephalography (EEG) signal was larger for errors than correct actions (Falkenstein et al., 1991; Gehring et al., 1993). The researchers accordingly named the component error negativity (Ne; Falkenstein et al., 1991) or *error-related negativity* (ERN; Gehring et al., 1993). Although the name error-related negativity might be misleading (see Alexander & Brown, 2011), the term ERN is well established and I will consequently use it in the current work to facilitate comparison to previous studies. The amplitude difference between ERPs after correct responses and errors is largest at frontocentral electrode sites, right above the mPFC, and evidence quickly accumulated linking it to mPFC and especially ACC activity (Debener et al., 2005; Dehaene et al., 1994; Kiehl et al., 2000; van Veen & Carter, 2002; for a review, see Taylor et al., 2007). In addition, the ERN latency shows that error processing happens fast; the ERN has its onset at or even before an erroneous action and peaks around 80 ms to 100 ms after it (Falkenstein et al., 1991; Gehring et al., 1993; for a review, see Gehring et al., 2012). The role of the mPFC and respective ERP components has been extensively studied over the past 40 years. I will begin by giving an overview of the neural and behavioral components underlying research on action monitoring before proceeding to theories on the precise functions of action monitoring in the brain.

## Neural and behavioral components of action monitoring

### *Own actions*

Going back to our example, let us imagine the acrobat in the air commits an error by flipping only half the way. At the moment of the error, her brain reacts, noticing the error and recalculating the next moves.

The ERN, as described above (Falkenstein et al., 1991; Falkenstein et al., 2000; Gehring et al., 1993; for a review, see Gehring et al., 2012), probably originates in the ACC (Dehaene et al., 1994; Ridderinkhof, Ullsperger, et al., 2004) and might be interpreted as the electrophysiological correlate of mPFC involvement in action monitoring. The ERP component occurs across different stimulus modalities (Forster & Pavone, 2008; Neta et al., 2015) and response modalities (Holroyd et al., 1998; Neta et al., 2015; Reinhart et al., 2012; van ’t Ent & Apkarian, 1999). Nevertheless, error responses might still slightly differ in latency, length, and topography, depending on task or response modality (Reinhart et al., 2012).

In addition to a brain response, errors elicit a robust control-enhancing behavioral response that is connected to processes in the mPFC (Danielmeier et al., 2011; Fu et al., 2019): errors often lead to an increase in response time in the trials following the error, called *post-error slowing* (Rabbitt, 1966, 1969). This effect is especially pronounced in speeded-response tasks (Buzzell et al., 2017; Danielmeier et al., 2011; Hajcak & Simons, 2008). Post-error slowing, as the ERN, presumably represents a control process (Kalfaoğlu & Stafford, 2014). However, although ERN amplitude and post-error slowing have been shown to correlate (see, for example, Debener et al., 2005; Gehring et al., 1993), dissociations have been found as well (e.g. Chang et al., 2014; Hajcak et al., 2003; Jentzsch et al., 2014). These mixed findings can be explained by specific factors modulating ERN and post-error slowing differentially, such as error awareness (Nieuwenhuis et al., 2001) or expertise (Jentzsch et al., 2014). As for the ERN, the exact mechanisms leading to post-error slowing are still debated. Some studies suggest that post-error slowing is caused by a change in the speed-accuracy-tradeoff as a result of errors (Marco-Pallarés et al., 2008; Rabbitt, 1966). More current research hints at a reorienting process after errors or unexpected events (Buzzell et al., 2017; Hajcak & Simons, 2008; Houtman et al., 2012; Jentzsch & Leuthold, 2006; Notebaert et al., 2009; Núñez Castellar et al., 2010). This theory states that (infrequent) error events lead to increased attention on these events and participants must reorient attention to the task. In line with this explanation, participants show no post-error slowing or even post-error speeding when responses are allowed to be slow, especially for difficult tasks (Damaso et al., 2020; Williams et al., 2016).

***Observed actions***

After the acrobat has made her error, the catchers must react very fast, adapting their position to safely catch the flyer. Processing observed errors, in fact, might be as important as processing own errors in our social environment. Even if (unlike or example) action consequences for observed actions might not always be immediate, observation offers the bonus that we can learn from others' mistakes. The brain activity of the catchers observing the error appears to be very similar to the acrobat's own response. Enhanced activity in ACC and mPFC when observing errors (de Bruijn et al., 2009; Shane et al., 2008; Yoshida et al., 2012, for a review, see Koban & Pourtois, 2014) suggest that similar monitoring mechanisms as in own actions take place. Additionally, observers display activity in pre-SMA and SMA (Scangos et al., 2013; Shane et al., 2008), which suggests that their brains simulate the motor responses that are observed, possibly through mirror neurons (Di Pellegrino et al., 1992; Gallese et al., 1996).

Results from EEG studies support the notion that own and observed actions are processed by the same system. For observed errors, a negative-going ERP component at frontocentral sites has been observed (Bates et al., 2005; Carp et al., 2009; de Bruijn & von Rhein, 2012; Miltner et al., 2004; van Schie et al., 2004). This component has sometimes been dubbed *observer error-related negativity* or error-related negativity for observation (Bates et al., 2005; van Schie et al., 2004). As the ERN, evidence suggests that the ERP component is generated in the mPFC, specifically the ACC (Miltner et al., 2004; van Schie et al., 2004). Because (unlike for the ERN) there is no well-established term for the component and its functionality is yet undetermined (Alexander & Brown, 2011; Bellebaum et al., 2020; Desmet et al., 2014), I will refer to the ERP component as *observer mediofrontal negativity* (oMN).

There are also differences between processing of own and observed actions. Only the processing of observed actions – not of own actions -, activates the superior temporal sulcus (Ninomiya et al., 2018), anterior insula (Cracco et al., 2016) and inferior parietal cortex (Shane et al., 2008). The oMN is usually smaller in amplitude than the ERN (Bates et al., 2005; Miltner et al., 2004; van Schie et al., 2004) and it usually peaks later (Miltner et al., 2004; van Schie et al., 2004). However, the oMN latency seems to strongly depend on the task. Peaks as early as 130 ms have been observed in Go/NoGo-Tasks (Bates et al., 2005; Koban et al., 2010), but for Flanker tasks, the oMN peaks 150 to 300 ms after the observed action (Carp et al., 2009; de Bruijn & von Rhein, 2012; Miltner et al., 2004; van Schie et al., 2004).

Behaviorally, post-error slowing effects have been observed for participants watching others make an error (Ceccarini & Castiello, 2019; Núñez Castellar et al., 2011; Schuch & Tipper, 2007). Processes of reorienting after own errors thus possibly extend to observed errors.

***Own and observed feedback processing***

The outcome of actions in terms of their accuracy can be determined using two types of signals. First, internal signals can be used that directly reflect our actions, and second, external signals can be used, such as feedback that we receive after the action. The acrobat might notice directly that her movements are not as expected, but she might also receive feedback on her action from her coach – or simply from the position in which she lands.

Feedback processing in the brain greatly resembles error processing. The mPFC, especially ACC, plays a major part in both (Becker et al., 2014; Holroyd et al., 2004). Negative feedback (compared to positive feedback) leads to a relatively negative ERP component around 250 ms after the feedback. This component was originally called *feedback-related negativity* (FRN; Hajcak et al., 2006; Miltner et al., 1997; Yeung et al., 2005). It supposedly originates in the mPFC (Gehring & Willoughby, 2002; Nieuwenhuis et al., 2004). The component seems to be connected to the reward circuit (Becker et al., 2014, see Shohamy et al., 2008), including ventral striatum and ACC and neutral outcomes elicit similar negative amplitudes as negative outcomes (Holroyd et al., 2006; Kujawa et al., 2013). As a consequence, recent theories propose that the difference in FRN amplitudes between positive and negative outcomes stems from a relative positivity after positive (rewarding) outcomes. Negative and neutral outcomes, according to these theories, lead to a baseline activation in form of an ERP negativity. In this light, researchers suggest to use the term *reward positivity* (rewP) instead of FRN for the difference signal between positive and negative outcomes (Proudfit, 2015). Consequently, I will refer to the component as rewP for the difference signal between positive and negative outcomes, and as FRN when referring to the negativity following feedback that is presumably reduced after positive outcomes.

As for error processing, a similar brain response as for feedback to own actions could be found in the catchers who observe the acrobat getting feedback from the coach. An ERP component similar to the rewP is also elicited, though with reduced amplitude and less consistently, when observing feedback given to others (Bellebaum et al., 2010; Fukushima & Hiraki, 2006, 2009; Itagaki & Katayama, 2008; Kang et al., 2010; Koban et al., 2012; Yu & Zhou, 2006). In a functional magnetic resonance imaging (fMRI) study, an involvement of the mPFC, specifically the ACC, has been found for observing feedback (Mobbs et al., 2009). However, the responses to observed feedback and the so-called *observer FRN* (oFRN) are strongly modulated by social context (Fukushima & Hiraki, 2009; Itagaki & Katayama, 2008; Kang et al., 2010; Koban et al., 2012) and interindividual differences such as gender and empathy (Fukushima & Hiraki, 2006, 2009; Koban et al., 2012). These findings raise the question whether oFRN and rewP might not be functionally separate, but a more general correlate of outcome evaluation, regardless of the actor (self or other) of the preceding action (Gehring & Willoughby, 2002; Yeung et al., 2005).

**Theories on action monitoring in the mPFC**

As established in the previous paragraphs, own and observed actions and feedback elicit brain responses that robustly vary depending on their valence. Errors and negative feedback result in increased mPFC activity and enhanced ERP component amplitudes. But how can the brain know that an error has happened? What are characteristics of actions and contexts that need to be met for the mPFC to be active?

The early research on error processing (Falkenstein et al., 1991; Gehring et al., 1993) also stated one of the first theories concerning the role of the ACC. They suggested that the region receives efference copies of the response and compares these to the representation of the intended response (Falkenstein et al., 1991; Falkenstein et al., 2000). In our example, the acrobat would compare her movement to the correct movement she has learned. If both are not the same, an ERN would emerge. If that theory held true, so Falkenstein et al. (2000) argued, the ERN should be larger the more the actual motion deviated from the aspired motion (supported by Bernstein et al., 1995; Falkenstein et al., 2000). Also, error responses should only show when an efference copy is created; the theory thus assumes that errors based on faulty knowledge elicit no ERN (Dehaene et al., 1994, supported by Scheffers & Coles, 2000; Tucker et al., 1993).

As a conflicting result to this early theory, Carter et al. (1998) found increased ACC Blood-Oxygenation-Level Dependent (BOLD) activity not only for incorrect responses, but also for correct responses when two possible responses competed. In our example, the acrobat has a whole repertoire of movements she can perform in the air, which compete with each other. If the conflict between them is high, activity in the ACC is enhanced. In Carter et al.'s *Conflict Monitoring Hypothesis*, they proposed that ACC activity (and subsequently the ERN) codes response conflict rather than errors per se (Botvinick et al., 2001; Carter et al., 1998; Yeung et al., 2004). Errors, according to the theory, are just a byproduct of the conflict – if conflict between the correct and erroneous response is especially high, the probability increases that the erroneous response is executed. The theory matched findings of increased ACC activity in tasks that enhanced conflict (Bench et al., 1993; Thompson-Schill et al., 1997) and could explain why unconscious errors could elicit an ERN. Van Veen and Carter (2002; see also Mathalon et al., 2003; Nieuwenhuis et al., 2003, for similar findings) showed a negative component in the ERPs originating from the ACC for high conflict correct trials before the response, which was similar, but earlier, as an ERN. They suggested that both negative components represent a conflict peak. The earlier the conflict, the more likely it is that a correct action occurs. However, more rostral areas of the mPFC were shown to be active only in high-conflict errors, and more dorsal areas in low-conflict errors, with overlap in the MCC (Wittfoth et al., 2008). Thus, while conflict might influence error monitoring, error processing cannot be fully explained by increased conflict in error responses. In addition, the Conflict Monitoring Hypothesis predicts larger ERNs for more frequent as opposed to less frequent errors (conflict between the responses is higher). However, the opposite pattern has been observed (see Holroyd & Coles, 2002).

In an influential theory proposed shortly after the Conflict Monitoring Hypothesis, Holroyd and Coles (2002) account for error-related activity in the ACC based on principles of reinforcement learning. The mesencephalic dopamine system, including the substantia nigra, basal ganglia and ACC, is fundamental in learning from rewards. Single dopamine neurons in the midbrain of monkeys show high activity for unexpected rewards, baseline activity for expected rewards and decreased activity for unexpected non-rewards (Hollerman & Schultz, 1998; Schultz et al., 1997). This indicates that the mesencephalic dopamine system codes expectations. Holroyd and Coles extended these findings to action monitoring in their *Reinforcement Learning Theory*. The theory predicts dopamine-related ACC activity for bad behavioral outcomes, in accordance with previous theories, but only if they are worse than expected. If the acrobat tends to make the same error over and over again, ACC activity should be smaller than for an error she makes for the first time. The Reinforcement Learning Theory assumes that *signed prediction errors* are reflected in ACC activity. The theory thus assumes a difference for unexpected negative actions or outcomes and positive actions or outcomes, but no difference between expected negative and positive actions or outcomes. Holroyd and Coles could show, accordingly, that infrequent errors elicited higher ERN amplitudes than more frequent errors. In addition, the theory can be applied to the relation between ERN and RewP. Holroyd and Coles could show that if participants could learn to predict the feedback based on their responses, the feedback-locked RewP (quantified as the difference between correct and incorrect, so technically an inverted RewP) de- and the response-locked ERN increased. This is much the same pattern as observed in monkeys' dopamine neurons during learning (Hollerman & Schultz, 1998; Schultz et al., 1997). In the following years, evidence accumulated both supporting the influence of expectancy on the ERN and RewP (Brown & Braver, 2005; Eppinger et al., 2008) and the connection between ERN and RewP (Bellebaum & Colosio, 2014; Eppinger et al., 2008; Pietschmann et al., 2008). However, no respective connection was shown during observation (Bellebaum & Colosio, 2014).

Opposed to assumptions made in the Reinforcement Learning Theory, ERN amplitudes and mPFC activity are increased not only for outcomes and actions that are *worse* than expected, but also for outcomes and actions that are *better* than expected (Ferdinand et al., 2012; Jessup et al., 2010; Núñez Castellar et al., 2010; Oliveira et al., 2007). The mPFC might thus code *unsigned* instead of *signed* prediction errors. Building on Holroyd & Coles' (2002) assumptions, Alexander and

Brown (2011) proposed the *predicted response-outcome* (PRO) *model*: they assume that the mPFC generates predictions on both actions and outcomes. If these predictions match the actual response or outcome, the mPFC response is inhibited. If the predictions do not match, mPFC activity is, respectively, higher. This interprets error (and other) activity of the mPFC as unsigned prediction error activity. For the acrobat and her observers, the PRO model would assume increased mPFC activity also if the acrobat performs a correct movement after she has made the wrong movement several times in a row, not only if she makes an unexpected error. The PRO model is able to explain almost all characteristics of mPFC and ACC activation: First, it explains its sensitivity to error likelihood, and the decrease of rewP amplitudes during learning (see Bellebaum & Colosio, 2014; Eppinger et al., 2008; Pietschmann et al., 2008; due to the predictability of both positive and negative feedback for a well-learned task). Second, it also explains conflict effects – the higher the conflict, the more difficult it presumably is to make predictions concerning action or outcome, and the less expected specific actions and outcomes are.

**Current Research on the PRO model**

The PRO model interprets brain responses following both own and observed actions or outcomes as a function of the unexpectedness of the event. The acrobat, therefore, generates a brain response when a movement or an external signal differs from what she expects. This might even be more relevant in the moment than defining the action’s valence; if the acrobat is able to predict the movement, erroneous or not, she can more easily adapt to the action and its consequences. The same holds for the observing catchers. They need to notice whether the acrobat lands somewhere unexpected, but they do not need to know why she lands there to catch her safely.

In recent years, more and more evidence has accumulated in support of the PRO model. As the model and the corresponding empirical findings build the main basis for the present work, the respective studies are described in detail in the following paragraph.

***Own Action Monitoring***

Several studies have been conducted on the influence of expectancies in own response monitoring. Wessel et al. (2012) showed that both the ERN and the novelty-related frontocentral N2 are generated in the mPFC, specifically in the pMFC. In a subsequent fMRI experiment, they observed comparable activity in the aMCC after errors and novel events. Patients with lesions to the left prefrontal cortex showed significantly reduced ERN and novelty-related N2 amplitudes compared to healthy controls (Wessel et al., 2014). They also did not show, as the control group, slowing after novelty or error (Wessel et al., 2014). The ERN and N2 thus seem to code similar processes, which might be related to expectancy. Both novelty events and (infrequent) errors reflect unexpected events, and error as well as novelty processing in the mPFC could be attributed to these expectancy violations. Núñez Castellar et al. (2010) used an adaptive algorithm that manipulated task difficulty. With this, they ensured predetermined error frequencies per participant, which should manipulate the expectancies of errors. The authors found a difference after correct and erroneous responses in ERN amplitudes only in the condition where errors were less frequent than correct responses (75% correct). They found no such difference in the condition where error responses were more frequent (35% correct). Gawlowska et al. (2018) used yet another expectancy manipulation. They employed a learning task where participants would naturally expect errors at the beginning of learning, with decreasing error expectancies as learning continued. They found that the processing of erroneous and correct responses as a function of ERN amplitude and ACC BOLD response did not differ at the beginning of the learning task. During learning, however, BOLD activations and ERN amplitudes after erroneous actions increased.

Similar effects as for action processing have been observed for feedback processing. Most of these studies used feedback type frequency to manipulate expectancies: Jessup et al. (2010) showed that ACC activity in response to negative feedback did not only decrease when negative feedback was more frequent, but positive infrequent feedback resulted in significantly more ACC activity than negative frequent feedback. Chase et al. (2011) found highest FRN amplitudes for infrequent negative feedback, and no significant difference between frequent positive feedback and frequent negative feedback. While these results suggest expectancy effects on ACC activity, they support signed prediction error activity (as proposed by Holroyd & Coles, 2002) rather than unsigned prediction error activity in the ACC. Ferdinand et al. (2012) constructed a time-estimation paradigm. Based on how much the participants deviated from the correct timing, they received negative (20%; high deviance from correct time point), positive (20%; low deviance from correct time point) or intermediate feedback (60%; intermediate deviance from correct time point). The thresholds for the respective feedback types were set adaptively to obtain the predetermined frequency. FRN amplitudes reflected frequency, but not valence. In contrast to these results, Núñez Castellar et al. (2010) observed no frequency effects on the FRN, possibly because of different predetermined feedback type frequencies (75% or 35% correct).

Instead of a frequency modulation, Oliveira et al. (2007) prompted outcome expectancies before giving feedback to the participants. In the task they used, errors, and subsequently, negative feedback, happened quite often (just under 50% of trials). The authors found lower FRN amplitudes after outcome-expectancy mismatches (participants stated that they had answered correctly and received negative feedback, or participants stated they had answered incorrectly and received positive feedback) than after matches. There was no effect of feedback valence. Interestingly, the study by Oliveira et al. also found that participants overestimated their own performance (i.e., the actual frequency of positive and negative feedback). This suggests that outcome frequencies cannot be exactly translated to participants' expectancies. In their deterministic learning task, Gawlowska et al. (2018) found that the difference between FRN amplitudes after positive or negative feedback increased during learning. This contradicts findings on decreased differences in FRN amplitudes between positive and negative feedback for other deterministic conditions (Eppinger et al., 2008). However, Eppinger et al. (2008) varied contingencies as a within-subject factor, while participants in the study by Gawlowska et al. always learned with 100% contingency. In this study, outcomes should become more expected, but negative feedback also became overall less frequent during learning. The comparison between the studies by Gawlowska et al. and Eppinger et al. shows that outcome contingencies, the interplay between different within-subject conditions, and frequency might influence expectancies. It can therefore be suspected that frequency modulations are not the only determinants of outcome expectancy, which I will address in further detail below.

Finally, effects of expectancy – again manipulated by response type frequency – have been found for behavioral responses to errors. Notebaert et al. (2009) observed post-event slowing after infrequent correct and wrong responses with deterministic feedback. They also showed post-event slowing after unexpected events. The results were confirmed in a subsequent study (Núñez Castellar et al., 2010). Houtman et al. (2012) found enhanced post-error slowing after infrequent compared to frequent errors irrespective of whether participants received feedback or not.

***Observed Action Monitoring***

The similarity between own and observed action monitoring would suggest that respective effects of expectancy can be transferred to observed action monitoring. Desmet et al. (2014) created a paradigm in which participants watched persons using a device. They could either use it correctly or incorrectly, and the outcome could be positive or negative (but only negative outcomes were possible for incorrect use). Additionally, in some trials a random picture (unexpected and unrelated to the task) appeared on the screen during the outcome presentation (creating novelty events similar to Wessel et al., 2012). Unexpected events and unexpected negative feedback elicited highest mPFC activation, followed by action errors and expected negative feedback. Correct actions and correct outcome elicited the lowest activation.

Again, most studies on the effect of expectancy in observed action monitoring manipulated expectancy with frequency: Schiffer et al. (2014) found increased BOLD signals in the dmPFC after videos of correct or incorrect knot-tying if actions of opposite valence had been seen in a preceding training session (compared to videos with matching valences between training and test session). Wang et al. (2015) found no difference in oMN amplitudes after correct and erroneous observed responses, if observed errors were frequent (80%, indicating a signed prediction error). However, they found a significant difference when observed errors and correct responses were equally frequent (50%). In contrast, Pezzetta et al. (2018) observed significantly higher ERN amplitudes after errors compared to correct actions in a virtual reality setting where errors were very frequent (70%). This study might be another example that frequency modulation does not necessarily equal expectancy modulation. Since Pezzetta et al. used grasp movements as the object of observation, which is a task often seen in everyday situations, the general expectation of erroneous actions might be low even if they were more probable in the experimental task. To identify effects of expectancy detached from frequencies, Kobza and Bellebaum (2013) and Bellebaum et al. (2020) used a false-belief paradigm in which instructions were used to shape expectancies, with constant event frequencies across conditions. They showed that oMN amplitudes after unexpected (correct or erroneous) observed actions were significantly higher than after expected (correct or erroneous) actions, respectively. This activation was shaped by trait empathy and oMN amplitude differences were observed only in highly empathic participants (Bellebaum et al., 2020).

Expectancy – or frequency – effects have also been shown for observed feedback. Kobza et al. (2011) manipulated observed feedback frequencies. They found a difference in oFRN amplitude for negative and positive feedback only when negative outcome frequencies were low. They also found a main effect of outcome frequency: higher oFRN amplitudes emerged for feedback associated with a low frequency of correct outcomes than for feedback associated with a high frequency of correct outcomes.

Finally, Wang et al. (2015) used error frequency to manipulate expectancy in an observational task and, in contrast to previous findings, found no effect of error frequency on post-error slowing after observational errors, but a main effect of observed action valence.

***Discussion of empirical findings on the PRO model***

Although the presented studies form a strong case in favor of the PRO model, some potential problems arise. In most studies, action or outcome expectancies are manipulated via event type frequencies, often without checking if the manipulation worked. In this sense, it is important to note that while event frequency influences expectancies of these events to some extent, the two concepts are not equal. Significant valence effects for the oMN and for the FRN have still been observed when both events were equally frequent (Wang et al., 2015; Yeung et al., 2005). Additionally, one study in which expectancies were directly measured (Oliveira et al., 2007) found an overoptimistic expectancy bias: participants expected to receive positive feedback more often than they actually did. Frequency modulations also do not consider baseline expectancies that might play a role in expectancy formation and might thus only work for previously unknown actions. As an example in observed action monitoring, Pezzetta et al. (2018) let participants observe grasp actions, a familiar task, and found no effect of error frequency on the oMN. In contrast, Schiffer et al. (2014) had participants observe knot tying, which was a novel task to the participants, and subsequently found an error frequency effect in mPFC activation.

The problem of manipulating expectancies seems to be particularly relevant for research conducted on observed action monitoring. In two studies, no effects of frequencies on observed action monitoring (Pezzetta et al., 2018) or post-event slowing (Wang et al., 2015) were observed. These and other studies vary error frequency to manipulate expectancies of errors (Pezzetta et al., 2018; Schiffer et al., 2014; Wang et al., 2015), but, as explained above, the effects of probability on actual expectancies can be influenced by biases (Oliveira et al., 2007). In an attempt to exclude frequency effects, Bellebaum et al. (2020) and Kobza and

Bellebaum (2013) modified expectancies only via instructions while keeping action probabilities equal across conditions. However, this paradigm comes with its own limitations. The authors used a false-belief task, meaning that observed correct responses in the false-belief condition were errors from the perspective of the observed player: The player answers incorrectly based on the information available to them, but for the observer, who is privy to additional information, the action is an error. In this respect, errors in the false-belief condition could be perceived by the observer as vicariously correct actions, and vice versa for correct responses in the false-belief condition. Based on vicarious error processing, the oMN found by the authors would still encode valence, not expectancy. This interpretation seems more probable as electrophysiological responses were dependent on trait empathy (Bellebaum et al., 2020).

The presented studies leave open two questions to investigate in observed error processing. First, can expectancy effects on observed action monitoring be found in a paradigm that rules out both, biases related to frequency modulations and vicarious error processing (i.e., perspective taking)? And second, are there possible modulators of expectancy formation that need to be accounted for in (observed) action monitoring research? Regarding the last point, the results of Bellebaum et al. (2020), if they can be attributed to expectancies, suggest a modulation of observed action monitoring by empathy as well as expectancy. The potential modulating effect of empathy on expectancy formation might be further investigated in follow-up studies.

Further contradictory findings concerning the PRO model have been published by Maier and Steinhauser (2016). In a flanker task, they varied both error frequency and error type. They used two error types, namely single errors (participants failed to react to the target but successfully ignored the flankers) and double errors (participants failed to react to the target and chose the flanker response). The authors found that double errors led to higher ERN amplitudes than single errors even if frequencies for both error types were equally high. When double errors were less frequent than single errors, but error type could be determined by the participants only at a later point (flankers occurred 100 ms after the target), ERN amplitudes did not differ. In compliance with the fact that even in studies where expectancies (or rather, frequencies), were accounted for, effects of valence were sometimes observed (Gawlowska et al., 2018; Jessup et al., 2010; Núñez Castellar et al., 2010; Pezzetta et al., 2018; Wang et al., 2015), this presents a third, albeit much broader, open question: are action monitoring processes for own and observed actions only dependent on expectancies, or are there some scenarios in which valence itself is (also) coded in the respective brain responses?

## Influences on Action Monitoring

The theories discussed above try to answer the question which mechanisms are responsible for the increased activity in action monitoring brain areas such as the mPFC and respective ERP components after errors/negative outcomes compared to correct actions/positive outcomes. Based on the PRO model (Alexander & Brown, 2011), several studies have shown a strong effect of expectancy on action monitoring. However, as mentioned above, additional factors can influence action monitoring and might interact with expectancies, such as trait empathy (Bellebaum et al., 2020; Lockwood et al., 2015), or show effects above the effect of expectancies, such as error type (Maier & Steinhauser, 2016). I will discuss two potential modulators of action monitoring, namely interindividual differences, and the (subjective and objective) significance of errors.

### *Interindividual differences*

The neural and behavioral correlates of action monitoring differ greatly between individuals based on several variables. The ERN is even discussed as a biomarker for psychiatric diseases due to systematic differences in ERN amplitudes between patients and healthy control groups. Enhanced ERN amplitudes compared to healthy controls were found in patients with obsessive compulsive disorder (Carrasco, Hong, et al., 2013; Endrass et al., 2008; Gehring et al., 2000) as well as their siblings (Carrasco, Harbin, et al., 2013). Enhanced ERN amplitudes were also observed for other anxiety disorders (Carrasco, Hong, et al., 2013; Endrass et al., 2014; Weinberg et al., 2010). For depression, the results are less coherent, with findings showing enhanced, decreased or non-altered amplitudes (Chiu & Deldin, 2007; Ladouceur et al., 2012; Ruchsow et al., 2004).

In contrast, ERN amplitudes are diminished in patients with schizophrenia compared to healthy controls (Bates et al., 2002; Morris et al., 2006) as well as in siblings of schizophrenia patients (Simmonite et al., 2012). Reduced ERNs were also observed in bipolar disorder patients (Morsel et al., 2014). Furthermore, addiction patients elicit reduced ERNs (Franken et al., 2007; Littel et al., 2012; Luijten et al., 2011), and this reduction can even serve as a relapse predictor (Marhe et al., 2013) and as an indication for risk groups (Euser et al., 2013).

In healthy participants, error monitoring seems to be linked, among others, to perfectionism (Barke et al., 2017; Perrone-McGovern et al., 2017; Stahl et al., 2015) and emotional intelligence (Perrone-McGovern et al., 2017). While the relationship between ERN amplitudes and, for example, obsessive compulsive disorder, is pretty robust (Carrasco, Hong, et al., 2013; Endrass et al., 2008; Endrass et al., 2014; Gehring et al., 2000), other mediating interindividual measures, e.g. empathy and expertise, are less well-understood.

*Empathy and Action Monitoring*

Empathy stands out as a specifically relevant modulator of observed action monitoring for two reasons. First, empathic reactions, in a neuroscientific sense, are defined by a similar brain activation pattern during the self-experience of certain emotional states (mostly pain) and observing emotional states in others (Bufalari et al., 2007; Lamm et al., 2011; Singer et al., 2004). As established above, action monitoring of own and observed actions also activates mostly the same brain regions (de Bruijn et al., 2009; Koban & Pourtois, 2014; Shane et al., 2008). This similarity between processing of own and observed emotional states and own and observed actions might suggest that observed action monitoring is an empathic response of the brain. Brain activation during pain observation is modulated by interindividual empathy differences (trait empathy, Singer et al., 2004; Singer et al., 2006) and by contexts that induce or reduce empathic feelings (state empathy; Singer et al., 2006). Similarly, observed action monitoring should be modulated by trait and state empathy. Second, and even more importantly, trait empathy might interact with effects of expectancy on observed action processing (Bellebaum et al., 2020; Lockwood et al., 2015) and, as a modulating factor, might account for some inconclusive results regarding the PRO model.

While no clear definition of empathy exists as of today, Bernhardt and Singer (2012) describe it as the ability of understanding and sharing others' emotions. Empathy is believed to contain both an affective and cognitive component (Davis, 1983). These subconcepts are supported by the different characteristics of psychiatric disorders associated with reduced empathy in either the cognitive (see autism spectrum disorder) or affective domain (see psychopathy; Jones et al., 2010). However, empathy is not a mere trait: The empathy felt towards another person as well as general empathic abilities can be influenced both by trait and state empathy (see Bernhardt & Singer, 2012). Trait and state empathy are closely connected, in that the one affects the other, and both affect empathic responses in the brain (Singer et al., 2004; Singer et al., 2006). Although I will focus on trait empathy in this work, I will also shortly discuss findings of state empathy on action monitoring. As mentioned above, state empathy is important to consider when interpreting trait empathy effects, because the latter might be influenced by the former.

In the context of empathy, it is important to distinguish between own and observed action monitoring. Observing actions is in itself a social process while own error monitoring can, but doesn't have to, happen in a social context. Consequently, we would expect stronger empathy effects on observed than own action monitoring. Still, one argument that might suggest a connection also between own action monitoring and empathy is that empathic feelings and abilities are closely connected to the ACC (Bernhardt & Singer, 2012; Lamm et al., 2010; Lamm et al., 2011; Singer et al., 2004). This could mean that this brain region influences both empathy and action monitoring, and respective cytoarchitectonic differences, such as grey- and white-matter density and interconnectivity of the ACC, might affect both processes.

Regarding *state* empathy, own action monitoring responses in the brain seem to be stronger in social contexts. Higher ERN amplitudes have been observed in an evaluation vs. a non-evaluation context (Hajcak et al., 2005), in competition vs. neutral context (van Meel & van Heijningen, 2010), and when a silent observer was present (Kim et al., 2005). However, an interaction with anxiety suggests that higher ERNs in social contexts might be due to a higher perceived significance of errors (Barker et al., 2015).

As for *trait* empathy, some studies indeed found higher ERN amplitudes for more empathic participants (Larson et al., 2010; Santesso & Segalowitz, 2009). However, a recent meta-analysis, which also includes the two aforementioned studies, found no relation between the two measures (Amiruddin et al., 2017). In conclusion, a relationship between own action monitoring and empathy seems improbable but has not been ruled out yet.

In action observation, a presumed effect of *state* empathy has been found in a number of studies. Observed action processing is modulated by the perceived similarity between own and observed person (Carp et al., 2009) and competitive vs. cooperative context (Koban et al., 2010). An effect of similarity on vicarious reward processing was shown in activation in ventral striatum and ACC (Mobbs et al., 2009). The more similar to themselves the observers rated the observed person, the stronger was the respective brain activation of the observer. Newman-Norlund et al. (2009) observed higher activation in the middle ACC when observing friends' vs. foes' errors. Similarly, Kang et al. (2010) observed larger oFRN amplitudes after observing friends than strangers, and Marco-Pallarés et al. (2010) even observed an oFRN after observed wins in a competitive context. However, de Bruijn and von Rhein (2012) found no difference between oMN amplitudes in a competitive vs. cooperative context.

With regard to *trait* empathy, studies are less conclusive. Brazil et al. (2011) found reduced amplitudes after both observed erroneous and correct actions for patients with psychopathy compared to healthy controls. The authors explained this with reduced trait empathy of the patients. Clawson et al. (2014), on the other hand, found no differences in the response to vicarious feedback between participants with autism spectrum disorder and healthy controls. Again, a possible effect might have been due to lower trait empathy in participants with autism spectrum disorder. Newman-Norlund et al. (2009) found that activation in the ventral ACC and pre-supplementary motor area after observed action errors correlated negatively with empathic concern. Empathic concern is a subscale of the interpersonal reactivity index (IRI; Davis, 1980, 1983) that is regarded as a part of affective empathy. Shane et al. (2009) showed that responses to observed errors in the inferior parietal cortex correlated with the IRI cognitive empathy subscale perspective taking, while activity in the ventral ACC correlated with the IRI affective empathy subscale empathic concern. Lockwood et al. (2015) found that the likelihood of vicarious rewards correlated with activity in a part of the aMCC, and that this correlation was modulated by trait emotion contagion. Emotion contagion is an affective empathy subscale of the Questionnaire of Cognitive and Affective Empathy (Reniers et al., 2011). Fukushima and Hiraki (2009) found an influence of fantasy (affective empathy subscale of the IRI) and emotional reactivity (affective empathy subscale of the Cambridge Behavior Scale, Baron-Cohen & Wheelwright, 2004), on electrophysiological measures of observed error processing only when real persons (as opposed to a computer) were observed. Finally, Bellebaum et al. (2020) found that a general empathy measure assessed with the Cambridge Behavior Scale (Baron-Cohen & Wheelwright, 2004), modulated ERP responses to observed unexpected and expected events.

In conclusion, while some effects of trait empathy on both own and observed error processing have been reported, the results are inconclusive as to when and which aspect of empathy is involved. In the case of own action processing, current research even suggests no connection between both processes. For observed error processing, other modulating factors seem to influence the effect of empathy on action monitoring. Both empathy and expectancy effects on observed action processing could be modulated by other processes, and an interacting effect on observed action monitoring has been suggested (Bellebaum et al., 2020; Lockwood et al., 2015). This leads to another open question: could inconsistencies in the findings regarding both the influence of expectancy and the influence of empathy be attributed to the interaction between the two factors?

*Expertise and Action Monitoring*

Another potential factor modulating both expectancies and action monitoring of own and observed actions might be expertise. Expertise describes the ability of a person to perform above average in a specific domain (see Ericsson et al., 1993), although it is difficult to determine at which point a person can be considered an expert. For acquiring expertise, both genetic (talent) and behavioral (practice) preconditions need to be met (Ericsson et al., 1993).

Regarding expectancies, expertise might facilitate predicting own and observed action and action outcomes (Özkan et al., 2019; Zhao et al., 2021): An experienced acrobat should be far better than a beginner at predicting, at any point in the stunt, what movements are likely to happen next or what potential outcomes a movement might have.

As for effects of expectancy on action monitoring, relatively few research has been conducted as of yet. Jentzsch et al. (2014) found that with increased musical expertise, participants exhibited larger ERN amplitudes in response to errors. Rachaveti et al. (2020) observed that post-error slowing after negative feedback decreased as a function of practice. Harris et al. (2014) found higher ERN amplitudes in a spelling decision task for participants with higher spelling abilities. However, at least the results of the last two mentioned studies might be confounded with error frequency, indicating the importance of combining expertise and expectancy measures in future studies. Error monitoring has been investigated in expert groups performing in their area of expertise such as skilled typists (Kalfaoğlu et al., 2018) and pianists (Herrojo Ruiz et al., 2009; Maidhof et al., 2009;

Paas et al., 2021), but these groups have not been directly compared to non-experts.

For observing actions, expertise in the observed action is beneficial, but not necessary. Panasiti et al. (2016) showed that motor expertise helps in detecting errors (shown in behavioral data), apprehending errors (shown in increased amplitudes in a positive ERP reflection following the ERN; see also Candidi et al., 2014 for motor-evoked potential data) and in motor simulation during observation (shown in left-lateralized mu suppression). However, Desmet et al. (2014) showed that activity in the mPFC was increased for errors even for actions that could not be performed by the observers themselves.

While the possible modulations of expertise on own and observed action monitoring pose an interesting research field, I mainly focus on the already-established effect of empathy in this dissertation. Especially for expert groups, however (as investigated in study 3), a potential effect of expertise on action monitoring – also as a modulator of other influences such as expectancy - is important to consider.

***Error Significance***

Action monitoring processes likely do not only differ between participants, but also depending on the specific action that has taken place. Specifically, research has indicated that the type of error influences action monitoring even if the error types are equally frequent and one should not be more expected than the other (Maier & Steinhauser, 2016).

In most studies mentioned above, actions have been coded as either right or wrong (or expected and unexpected). However, errors can be manifold: If the acrobat forgets to point her toes during the flip, it will look less pretty, but if she only flips half-way, it might be dangerous. Those different levels of errors are common in musical performance or sports. In addition, there is the subjective significance of an error. If the acrobat has a tendency to not point her toes and has difficulty correcting this, she might consider the error to be more important than another acrobat would. These different types of errors have been investigated only sparsely and in limited areas.

The subjective importance of an error seems to be relevant for action monitoring. Higher motivation to perform correctly has been shown to lead to larger ERN amplitudes, with modulation by outcome consequences (Ganushchak & Schiller, 2008; Hajcak et al., 2005) and personal traits such as perfectionism (Barke et al., 2017; Perrone-McGovern et al., 2017; Stahl et al., 2015). Similar results were obtained in observed action processing studies (Koban et al., 2010; Newman-Norlund et al., 2009). Another index for subjective error significance might be automatic error correction, and accordingly, automatically corrected errors are accompanied by higher ERN amplitudes (Paas et al., 2021) or earlier peaks (Fiehler et al., 2005).

Regarding error type, ERNs are larger for errors that require larger responses (under-reach vs. over-reach; Murata & Katayama, 2005). Especially the severity of an error might influence error processing (as severe errors would almost always equal subjectively more significant errors). Bernstein et al. (1995) conducted a stimulus-response study with four response options (two fingers on each hand), and compared single (wrong hand or wrong finger) to double errors (wrong hand and wrong finger). Double errors elicited significantly higher ERN amplitudes. In a Flanker task, Maier and colleagues (Maier et al., 2008; Maier et al., 2012; Maier & Steinhauser, 2016) also compared single (participants did not respond correctly to the target) and double errors (participants did not respond correctly to the target *and* failed to ignore the flankers) and found similar results. Importantly, Maier and Steinhauser (2016) determined that this pattern could not be attributed to error type frequencies. For both paradigms, it is still unclear if the higher ERN amplitude for double errors represents combined ERNs for both errors or if double errors are merely perceived as larger errors that then lead to increased amplitudes. For observed actions, small errors might be recognized less accurately than larger errors, and effects of severity on later ERP component such as P300 and N400 were shown (Amoruso et al., 2014). However, no studies investigating the effect of error severity on the oMN exist as of now.

**Addressing Open Questions**

As established above, the role of the mPFC as action monitoring instance has been shown in a great number of studies. However, the exact mechanisms and possible modulators of this process are still under investigation. For a start, the PRO model (Alexander & Brown, 2011) suggests that expectancies regarding actions or outcomes shape mPFC responses to own and observed actions and outcomes. Previous effects of action/outcome valence can be explained by the infrequency in which errors occur in most paradigms. While the model seems to fit most results in action monitoring studies, results on observed action monitoring do not yet show a definite picture. This is mainly due to conceptual problems with regard to the studies. Most of them used a modulation of event type frequencies to manipulate expectancies (see Pezzetta et al., 2018; Schiffer et al., 2014; Wang et al., 2015), which, however, might be confounded by previous task experience (Pezzetta et al., 2018) and an overoptimistic bias concerning the actual performance (Oliveira et al., 2007). Studies in which expectancies were manipulated with a false-belief condition, however (as in Bellebaum et al., 2020; Kobza & Bellebaum, 2013), pose the problem that an error in the false-belief condition might be perceived as a vicarious correct action. We address the question whether observed action monitoring reflects expectancies (not subjective evaluations of frequencies or vicarious error processing) in study 1. In addition, there is a possible modulating effect of empathy on the effects of expectancy on observed error processing (Bellebaum et al., 2020), and vice versa, an effect of expectancy on findings regarding empathy. This possible modulation calls for further investigation regarding the exact relationship between the two factors. A special focus on the exact empathy mechanisms involved in observed action monitoring is set in study 2, although empathy effects are investigated in all three studies. Finally, the effect of error severity on monitoring of both own and observed actions has been neglected in most previous work, also in its possible interaction with expectancies (see Maier & Steinhauser, 2016), empathy (mainly in observed actions) and expertise. This open question is investigated in study 3.

# Overview of Studies

## Study 1

In this study, we aimed to eliminate confounds of expectancy induction in observed action monitoring. We used a scenario in which expectancies were induced not only by a false-belief condition, but also by varying the perceived difficulty of the observed task, while error frequencies were held constant. The focus was on a negative frontocentral ERP component previously associated with error observation (called oMN in the following summary). We assessed behavioral measures of expectancy as potential predictors of the oMN amplitude as well as trait empathy which may play a modulating role.

### *Introduction*

As established in the General Introduction, action monitoring for observed actions elicits a similar electrophysiological ERP component as own action monitoring (van Schie et al., 2004, see also Koban & Pourtois, 2014). Both components are presumed to originate in the mPFC, probably the ACC (Dehaene et al., 1994; Taylor et al., 2007 for the ERN; Miltner et al., 2004; van Schie et al., 2004 for the oMN). However, the ERP component for observed action is smaller and peaks later, usually 100 to 300 ms after the observed error, although latencies might depend on the paradigm (Bates et al., 2005; Koban et al., 2010 as opposed to Miltner et al., 2004; van Schie et al., 2004).

Recent evidence supports the assumption that ERP components related to own action monitoring reflect unexpectedness rather than valence (Alexander & Brown, 2011; Ferdinand et al., 2012; Wessel et al., 2012). This extends to observed action monitoring: action prediction errors, i.e., the degree to which an observed action is unexpected (Burke et al., 2010; Donnarumma et al., 2017; Flanagan & Johansson, 2003), are assumed to shape oMN amplitudes. Kobza and Bellebaum (2013) used a false-belief paradigm to investigate expectancy effects on the oMN. In the false-belief condition, where errors were expected, they found larger negative amplitudes after correct than erroneous responses. However, the players' objectively correct answer in the false-belief condition was an error from the players' perspective; it only resulted to be correct because circumstances were changed. Thus, the oMN found in the study could have been shaped by vicarious error processing instead of expectancies. This is further supported by the effect of empathy (measured with the Cambridge Behavior Scale; Baron-Cohen & Wheelwright, 2004) on action processing in the task (Bellebaum et al., 2020). The authors replicated the effect found by Kobza and Bellebaum, but only for highly empathic participants, while lowly empathic participants did not differ in their oMN response between conditions. Therefore, the open questions remain whether observed action processing is a function of expectancy or vicarious error processing, and whether effects of empathy can be found if subjective action valence is accounted for.

To investigate these questions, we used the same false-belief paradigm as used by Kobza and Bellebaum (2013) and Bellebaum et al. (2020), but added the factor difficulty: errors in difficult trials are errors from the observed person's *and* observer's perspective. Meanwhile, errors should still be more expected in difficult than in easy trials because difficulty leads to a reduction in performance.
We expected that the amplitude of the oMN should reflect expectancy, not vicarious action valence. Additionally, we expected empathy to influence this process and aimed to further investigate the relationship between empathy and expectancy regarding observed action monitoring.

***Method***

We recruited 33 healthy participants to take part in the study. We used an adaptation of the false-belief paradigm introduced by Kobza and Bellebaum (2013; see also Bellebaum et al., 2020). In this task, participants observed another person playing a two-shell game. The game was explained to participants as follows: The observed player would see two shells, one of which was placed over a ball. Afterwards, the shells would rotate multiple times, and subsequently, the player would have to indicate via joystick what shell they believed the ball to be under. Observers were told that they were able to see the ball at all points, but the observed person could not. In accordance with previous studies (Bellebaum et al., 2020; Kobza & Bellebaum, 2013), we introduced a trick condition in half of the trials. In this, the ball was swapped between the two shells which was known to the observer, but not the observed person (false-belief). We also added the factor difficulty. In difficult trials, the shells were turned so fast that it seemed difficult for the observed player to follow, and participants were told that the player had to guess the correct answer. We therefore induced three levels of expectancies. Correct answers were expected in low-difficulty no-trick trials, guessing (so 50% correct) was expected in high-difficulty trick- and high-difficulty no-trick trials, and incorrect answers were expected in low-difficulty trick trials. In fact, the observed player answered correctly 50% of the time in all conditions to avoid effects of event frequency. We used a block design to vary between trick and no-trick trials because the swapping of the ball was hard to perceive in high-difficulty trials. By using a block design, observer participants knew beforehand if a trick or no trick would occur. There were four blocks in total (of 117 trials each), with alternatively all trick or all no-trick trials, and the starting condition was determined randomly to avoid order effects. The difficulty condition was modulated within blocks, with a randomly determined order. For each trial type (trick or no-trick), half of the trials were presented in the high difficulty condition, and half in the low difficulty condition. In 48 trials across blocks (12 each for the low difficulty no-trick, low difficulty trick, high difficulty no-trick and high difficulty trick condition) instead of observing a response, participants were asked where they thought the player would point to (prompt trials). These trials served both as a manipulation check and as dependent variable for later behavioral expectancy analyses.

We measured empathy with the German version of the Cambridge Behavior Scale (Baron-Cohen & Wheelwright, 2004; de Haen, n.d.) in accordance with Bellebaum et al. (2020). EEG signals at 30 passive electrodes (F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, CP3, CPz, CP4, P7, P3, Pz, P4, P8, PO7, PO3, POz, PO4, PO8) as well as EOG signals from eye electrodes were recorded.

We determined expectancies of participants in each trial condition (low/high difficulty x trick/no-trick) as the percentage of prompt trials in which they expected the player to choose the shell under which the ball was hidden. Subsequently, we performed a linear mixed effect (LME) analysis on expectancy with factors Trial Type, Difficulty and Empathy. Random slopes for Trial Type and Difficulty and random intercepts were allowed per participant.

For the EEG data preprocessing, a 0.5-Hz high-pass and a 20-Hz low-pass filter were applied to the recorded data. Independent component analysis and back-transformation was used to remove one component that represented blinks to reduce eye movement artifacts. Segments from -200 ms to 600 ms around the observed response were created and then baseline-corrected (the 200 ms pre-response served as baseline). An automatic artifact rejection was performed before creating averages per participant and condition (Trial Type x Difficulty x Accuracy) and exporting them for further analyses. In a visual inspection of the EEG data, we observed possible modulations at an earlier time window than in the previous studies (Bellebaum et al., 2020; Kobza & Bellebaum, 2013). As latencies of the oMN can differ depending on the task, we thus analyzed both, the earlier and later negative peak. The negative peaks were determined per participant and condition (Trial Type x Difficulty x Accuracy) on average ERPs in time windows of 100-250 ms and 250-420 ms, respectively. For both peaks, we determined the respective preceding positivities (in time windows from 50 ms to negative peak and 130 ms to negative peak, respectively) and subtracted them from the negative peaks to obtain peak-to-peak oMN amplitude values. We performed LME analyses with Trial Type, Difficulty, Accuracy and Empathy on both values. We allowed random slopes for Trial Type, Difficulty and Accuracy and random intercepts per participant.
In a post-hoc analysis, we determined the effect of behavioral expectancy values (per condition and participant) on the oMN amplitude, while allowing random intercepts per participant. We used model comparisons to check if empathy explained additional variance when added to the model.

***Results and Discussion***

We successfully verified that our manipulation worked. Expectancies for correct answers were highest in the low-difficulty no-trick condition and lowest in the low-difficulty trick condition. Expectancies were further modulated by empathy: we found stronger expectancies in both trick conditions for high empathic participants compared to low empathic participants in low-difficulty trials. A similar effect was not observed in the previous study, which entailed only low difficulty trials (Bellebaum et al., 2020). In this study, however, there was less variance between individual expectancies as in our task, suggesting a ceiling effect. We suspect that since our task included more factors, expectancy formation was more difficult and subsequently, there was more interindividual variation and the ceiling effect was reduced. This, in turn, allowed us to observe an effect of empathy on expectancies. Expectancies in no-trick conditions differed between high and low difficulty, but not in trick trials, possibly because of less strong expectancies in general in the trick trials.

Concerning the ERPs, we found modulations of expectancy only on the early ERP component. oMN latencies differ based on task (see Koban & Pourtois, 2014), and in the present study, due to the block design, very early expectancy formation (at the beginning of the shell-turning) was possible, which might have led to shorter latencies.

There was a four-way interaction of Empathy, Trial Type, Difficulty and Accuracy in the early ERP component. Resolving this, we found an effect of empathy (higher amplitudes for higher empathy values) for trick low-difficulty trials with correct response and no-trick low-difficulty trials with incorrect response. Additionally, we found no effect of empathy on high-difficulty trials, but higher amplitudes for correct answers in high difficulty no-trick trials and a trend for higher amplitudes for incorrect answers in high difficulty trick trials. In conclusion, the results suggest an effect of expectancy on the oMN, as the pattern of ERP amplitudes mostly mirrors the pattern of expectancies of observed actions in the different task conditions. A modulation based on (vicarious) valence processing can be excluded (in line with Bellebaum et al., 2020; Ferdinand et al., 2012; Kobza & Bellebaum, 2013; Schiffer et al., 2014), because such an effect should be irrespective of task difficulty. The pattern of behavioral results matches that of the early ERP peak for low-difficulty trials: Observed responses in trick trials are expected to be incorrect for high-empathic participants and observed correct responses in this condition elicit higher oMN amplitudes than observed incorrect responses for high-empathic participants. The opposite is true for high-empathic participants in no-trick trials. In contrast, the patterns are reversed in high difficulty trials. (Prediction) errors have been suggested to lead to increased attention on the source of the respective error (Notebaert et al., 2009; Steinhauser & Andersen, 2019; Wessel, 2018). Due to the block design in this study, low and high difficulty in each trial type were directly compared and thus, attention might have been paid to the comparison as opposed to absolute probabilities.

For the late ERP peak, we only found a modulation of accuracy in difficult trials. This suggests a functional difference between early and late observed error processing: processing of expectancies might happen early, followed by valence processing.

When Trial Type, Difficulty and Accuracy were replaced by individual expectancy values of the observers as predictors for the early ERP peak amplitude, a significant expectancy effect emerged. The more the respective answer was expected, the smaller were the (early ERP peak) amplitudes. Also, empathy did not explain significantly more variance when added to the model. We suspect that empathy might affect oMN amplitudes via a positive effect on expectancy formation, which is in line with our finding of an empathy effect on behavioral expectancy measures. As stated by Brown and Brüne (2012), prediction in non-social and social contexts might depend on similar processes, but for social contexts, some additional factors – such as empathy – might play a role (see also Fukushima & Hiraki, 2009). Depending on the specific context, specific aspects and empathic resources might be needed, which might explain inconclusive results in previous research.

***Conclusion***

Expectancies, not vicarious errors, are reflected in the oMN, although the effect had a shorter latency than expected. In our results, behavioral and electrophysiological dependent measures were modulated by empathy, but when behavioral expectancy values were used to predict electrophysiological responses, empathy did not explain any additional variance. We therefore assume that empathy modulates expectancy formation which in turn modulates neural action processing, and empathy is necessary for expectancy formation in some, but not all contexts.

**Study 2**

In study 2, we empirically tested the indirect effect of empathy on observed action monitoring (measured with the oMN) via expectancy formation. To do this, we calculated single-trial expectancy estimates and tested their effect on ERP amplitudes, then used model comparisons to identify any further effects of empathy.

***Introduction***

Recent findings suggest that the supposed valence effects in observed action processing are a function of the expectancy of events (Alexander & Brown, 2011): Observers experience an action prediction error if their predictions concerning others' actions are not met (Brown & Brüne, 2012; Burke et al., 2010). This influences frontocentral activity (Desmet et al., 2014; Schiffer et al., 2014) and leads to a negative ERP component (study 1; Bellebaum et al., 2020; Kobza & Bellebaum, 2013; Wang et al., 2015) that I will refer to as oMN in this summary. Action observation happens in a social context, so a link between empathy and action observation could be presumed. This has been shown for state empathy (Carp et al., 2009; de Bruijn & von Rhein, 2012; Kang et al., 2010; Koban et al., 2012), but results on trait empathy are inconclusive (Brazil et al., 2011; Clawson et al., 2014; Lockwood et al., 2015; Newman-Norlund et al., 2009; Shane et al., 2009). The inconclusive results could be explained by effects of prediction on action observation (Alexander & Brown, 2011; Bellebaum et al., 2020; Kobza & Bellebaum, 2013; Wang et al., 2015).

Empathy possibly facilitates expectancy formation in social situations (Brown & Brüne, 2012), but this might be task dependent, with, for example, higher need of empathy for false-belief tasks (see study 1; Bellebaum et al., 2020; Ferguson et al., 2015). In study 1, we assumed that expectancy formation might be facilitated by trait empathy and then might lead to respective processing of observed actions in the brain. While an effect of empathy on expectancy formation was found in this more complex task, no effect was found for a simple false-belief paradigm (Bellebaum et al., 2020). In the study by Bellebaum et al. (2020), expectancy induction was quite successful, with little variance, in both the true- and false-belief conditions of the task. A higher variability in the sample might be necessary to find potential empathy effects on expectancy. Also, the expectancy-inducing instructions differed from the actual event frequencies, and participants might have adapted their expectancies during the experiment. However, trial-by-trial variations were not accounted for in previous studies.

We therefore extended the sample of 50 women by Bellebaum et al. (2020) with a similar sample of men with varying backgrounds to increase variability in the induced expectancies. We also accounted for changes in expectancy over the experimental course by using single-trial expectancy values as predictors of single-trial oMN amplitudes. We hypothesized that expectancy formation would be modulated by empathy and that this effect would possibly change over the experimental course. In accordance with the post-hoc analysis performed in study 1, these expectancy values should predict frontocentral negative component amplitudes, but no further effect of empathy should emerge.

***Method***

The original sample consisted of 50 women (Bellebaum et al., 2020) and 55 men were added. As one woman and two men were excluded based on missing data, there were 102 participants in total. We employed the two-shell game used by Kobza and Bellebaum (2013) and Bellebaum et al. (2020) in which a ball was hidden under one of two shells which then rotated. Participants were told that the observed player watched this procedure and would subsequently choose (via joystick) under which one of the two shells they believed the ball to be. The observer participants always saw the ball and observed the game from above. In 50% of trials, the ball was swapped between shells, i.e., the player was tricked. Observer participants were told that this was visible to them, but not the player, making this a false-belief condition. Observer participants should expect correct answers in no-trick and wrong answers in trick trials. In fact, the player answered correctly 50% of the time in both conditions. In total, there were 420 observation trials, and false- and true-belief trials were presented in random order. In 48 additional trials (24 for each trick condition) trials ended not with the player answering, but with a question (prompt trial) to the observer: “where do you think the player will point the joystick?”.

We acquired the behavioral expectancy data from participants’ answers to these questions and assessed empathy with the German version of the Cambridge Behavior Scale (Baron-Cohen & Wheelwright, 2004; de Haen, n.d.) before the experiment. We recorded EEG with 29 active electrodes (F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO9, P1, Pz, P2, PO10 and FCz, which was used as online reference). We also recorded EEG at both mastoids to serve as offline reference and recorded additional EOG measures.

For the behavioral analysis, we calculated an LME model with expectancy (0 = observed response was expected to be incorrect, 100 = observed response was expected to be correct) as dependent single-trial variable, and the fixed effects Trial Type (trick/no-trick), Empathy, and Trial Number (to account for changes during the experiment). Both Empathy and Trial Number were continuous and mean-centered. Random slopes and intercepts per participant for Trial Type and Trial Number were included.

For the EEG data, we first re-referenced all data to the mean mastoid signal. We filtered the data with 0.5 Hz high-pass and a 20 Hz low-pass filter. ICA and ICA-back-transformation were used to remove one component representing blinks. We then segmented the data starting -200 ms pre- and ending 600 ms post-event. Afterwards, we performed baseline-correction and used automatic artifact rejection to exclude bad segments. We exported both single-trial data and averages by condition (Trial Type x Accuracy) and participant and pooled the data over a frontocentral electrode cluster: Fz, FC1, FCz, FC2 and Cz. In the average signals for each condition and participant, we determined the latency of the maximum negative peak between 250 and 420 ms, and the latency of the preceding positive peak between 130 ms and the negative peak. Then, we extracted single-trial values at these latencies, determined the difference between negative and positive peak for each trial and used this as a dependent variable to represent oMN peak-to-peak amplitudes.

We first aimed to replicate the results of Bellebaum et al. (2020) in a larger sample and with single-trial analysis. For this, we included factors Accuracy (correct or incorrect), Trial Type (trick or no trick) and Empathy in a model with peak-to-peak single-trial oMN amplitudes as dependent variable, and Accuracy and Trial Type as random factors by participant. In a second analysis, we first calculated two linear regression models of expectancy (the expectancy for a correct answer), for trick and no-trick, for each participant, determined based on the 24 data points for each trick condition and participant. The models allowed to estimate expectancy values for each trial. For the statistical analysis, we replaced Trial Type and Accuracy with the values of the respective regression model at the trial number. For observed error responses, we used inverted values, as the expectancy of an observed error was considered the reverse of the expectancy of a correct response, as which we coded participants' answers to the prompt trials. We then checked whether Accuracy, Trial Type or Empathy explained additional variance when added to the model.

***Results and Discussion***

The expectancy manipulation worked, meaning that expectancies of correct answers differed significantly between trick and no-trick trials, with lower expectancies of correct answers in the trick trials than in the no-trick trials, although expectancies were less strong and more variable than in the study by Bellebaum et al. (2020). In the behavioral LME model, we found a corresponding main effect of Trial Type, as well as a three-way interaction between Trial Type, Trial Number and Empathy. Expectancies grew less strong across the experiment only in trick trials for participants with low empathy. We found no overall effect of changing expectancies across trials although induced expectancies concerning error and correct responses differed from the actual frequencies of these events. This was possibly due to a confirmation bias (Nickerson, 1998; Talluri et al., 2018; Urai et al., 2019). As for the reduced strength of expectancies for low-empathic participants in trick trials towards the end of the experiment, we suspected that since trick trials (as the false-belief condition) require more empathy than no-trick trials, upholding expectancies might be especially hard for low-empathic participants, resulting in a fatigue effect. This is in line with studies suggesting that empathizing is reduced when cognitive load is high (Apperly et al., 2008; Epley et al., 2004) and that propose a social working memory system (Meyer et al., 2012) that might be impaired in low-empathic participants.

We were able to replicate the three-way interaction for the oMN found in Bellebaum et al. (2020). There was an interaction between Accuracy and Trial Type only for high empathy participants: the slope for Accuracy was inverted depending on Trial Type, with highest amplitudes for unexpected correct (in Trick Trials) and unexpected erroneous actions (in No-Trick Trials). With this analysis, we could show that single-trial data yield comparable results to analyses based on condition averages.

For our main analysis, we found a significant effect of Expectancy on single-trial peak-to-peak amplitudes of the oMN, and neither Trial Type, Accuracy nor Empathy explained significantly more variance when added to the model. We conclude that expectancy, not valence, influenced amplitudes (see Desmet et al., 2014; Ferdinand et al., 2012; Jessup et al., 2010; Oliveira et al., 2007; Schiffer et al., 2014; Wessel et al., 2012; Wessel et al., 2014; Zubarev & Parkkonen, 2018, for corresponding results). The results also support an indirect influence of empathy on amplitudes via expectancy formation (as also found in study 1). This could explain inconclusive results on the effect of empathy on action monitoring (e.g. Brazil et al., 2011; Clawson et al., 2014). The ACC could either contribute separately to both empathy and error monitoring, or different parts of the ACC might be at play. The study design had some limitations: first, using a binary variable in the regression model might lead to information loss, and single-trial expectancy values were only an approximation of the actual expectancies because they were based on only 24 values. Additionally, the single trial EEG data calculation, while it allowed to include more data than an average data approach, still includes only a fraction of the acquired EEG data.

***Conclusion***

Expectancy formation is dependent on empathy, and empathizing might not only be generally more difficult, but also more difficult to maintain over time for low-empathic participants. Expectancies directly shaped ERP responses, and no additional variance was explained by Accuracy or Empathy. In future studies, expectancies could be measured in every trial to obtain more accurate values.

**Study 3**

In study 3, we aimed to investigate another modulator of action monitoring besides expectancy, namely error significance. The PRO model (Alexander & Brown, 2011) cannot explain all effects of (subjective and objective) error significance on action monitoring (Ganushchak & Schiller, 2008; Hajcak et al., 2005; Maier & Steinhauser, 2016). At the same time, most studies investigating action valence in both, own and observed action monitoring, do so in a binary fashion, differentiating only between errors and correct actions. Therefore, we investigated whether action valence affects own *and* observed action monitoring (as seen in ERN and oMN, respectively) in a dichotomous or continuous way. We additionally investigated whether empathy or expertise modulated this effect and examined further effects of subjective error importance and the interplay between error severity and expectancy.

***Experiment 1***

*Introduction*

Research contrasting errors vs. correct actions has shown a clear pattern of activity in the mPFC (Debener et al., 2005; Ullsperger et al., 2014), corroborated by a negative mediofrontal component in electrophysiological data after errors, the ERN (see Gehring & Willoughby, 2002; Holroyd & Coles, 2002) While some aspects of error processing can be explained by the PRO model (Alexander & Brown, 2011), which suggests that the mPFC codes expectancies, valence, especially error significance, might still be coded in the mPFC. Action valence is often gradual, and behavioral adaptation (a function requiring the ACC, Devinsky et al., 1995; Holroyd & Coles, 2002) often needs to happen very quickly. Therefore, we suspect error severity to be represented in early action monitoring as coded in the ERN. In previous studies, double errors led to higher ERN amplitudes than single errors (Bernstein et al., 1995; Maier et al., 2012; Maier & Steinhauser, 2016). To address the role of error severity on own action monitoring, we conducted Experiment 1 with pianists playing while EEG was recorded. There were three types of events: correct responses, small errors (one note off), and large errors (two notes off).

The paradigm allowed to investigate behavioral variables, namely post-event reaction times and keystroke volume. Post-error slowing (for examples in piano tasks, see Herrojo Ruiz et al., 2009; Maidhof et al., 2009; Maidhof et al., 2013; Paas et al., 2021) possibly stems from an attentional shift towards the error (or unexpected event) and a successive attention reorienting back to the task (Notebaert et al., 2009; Núñez Castellar et al., 2010). Errors are associated with reduced keypress volume (Herrojo Ruiz et al., 2009; Maidhof et al., 2009; Maidhof et al., 2013; Paas et al., 2021), but the processes behind this are yet unclear.

We expected increased ERN amplitudes for large compared to small errors. Furthermore, the attention deviation should be stronger and thus, post-error slowing should be higher for large vs. small errors. We refrained from predicting error severity effects on keypress volume due to the lack of research on the underlying processes.

Musicians might have higher empathic abilities compared to other people (Gujing et al., 2019; Rabinowitch et al., 2012). An effect of trait empathy on ERN amplitudes is improbable (Amiruddin et al., 2017), but cannot be ruled out, so we included empathy as a possible modulator. In addition, pianists form an expert sample with training in error management (Kruse-Weber & Parncutt, 2014; Palmer & Drake, 1997). Different processing of errors depending on expertise has been shown (Jentzsch et al., 2014), so error processing might be modulated by expertise in our sample. With our paradigm, we could attempt to replicate the study by Paas et al. (2021). The authors investigated errors in a piano-playing task and found higher ERN amplitudes for automatically corrected compared to uncorrected errors. Finally, to investigate the results in light of the PRO model (Alexander & Brown, 2011), we performed additional checks if any effect of error severity might be explained by factors possibly influencing expectancies (Event Type Frequency, note Difficulty, and Insecurity while playing).

*Method*

We analyzed data from 21 pianist participants that all made at least 10 large errors. Six short pieces (played on only white keys with only the right hand) had to be learned two weeks before testing. The pieces were designed to provoke large errors. Participants filled out self-reports to measure Expertise and Empathy (Cambridge Behavior Score; Baron-Cohen & Wheelwright, 2004; de Haen, n.d.). After 29 active electrodes (FCz [which was used as online reference], F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO9, P1, Pz, P2 and PO10), as well as mastoid electrodes and EOG, were attached, participants played the pieces in random order 10 times each (so 60 sequences in total). Their performance was recorded on a laptop with a webcam placed above the piano to acclaim videos later used in Experiment 2. Participants were instructed to play slightly faster than during practice and to keep to their tempo even if it meant making errors. While playing, we recorded EEG (an EEG marker was sent every 5$^{th}$ keystroke) and musical digital interface (MIDI) data.

For the analysis, the MIDI data were compared offline to the correct score with a dynamic score-matcher algorithm (Large, 1993; Palmer & van de Sande, 1993; Rankin et al., 2009) to determine correct and erroneous keypresses. We calculated two behavioral variables that later served as dependent variables. First, the inter-keypress-interval (IKI) as the difference between onset of the current and next keypress, which served as a measure for post-event reaction times. Second, we obtained the volume of each keypress (this equals to keypress velocity in piano playing, which is recorded in the MIDI signal). Additionally, three measures potentially influencing expectancy were calculated to later serve as independent variables, namely Event Type Frequency (how often each event happened for each participant), Difficulty (how often each individual note was played correctly in the ten times it was played) and Insecurity (absolute deviation from mean velocity). From the behavioral accuracy data, we determined the Event Types correct keypress, small errors, and large errors (as well as the additional Event Type corrected small errors). We included only events that were preceded and followed by correct notes and that did not contain systematic errors (≥ 40% accuracy).

For the behavioral data LME analyses, we determined Event Type as fixed and random effect. IKI was set as dependent variable for the first model, volume as dependent variable for the second model. We checked for possible effects of Expertise and Empathy with model comparisons and, if any or both variables explained more variance, added them to the final model.

For the EEG data, markers were first recoded offline to match the behavioral data before EEG preprocessing was conducted. We used a 30 Hz low-pass and 0.5 high-pass filter. Subsequently, we employed the Gratton & Coles algorithm (Gratton et al., 1983) to remove blinks and eye movement artefacts. The data were segmented around each Event Type (300 ms before to 600 ms after keypress). An automatic artifact rejection removed bad segments. Baseline correction was performed on a baseline of -300 to -200 ms. Averages per participant and condition and single-trial data were exported for further analysis, and data were pooled over electrodes Fz, FCz and Cz. We extracted single-trial signals, that is, the mean amplitude in an area of -10 to +10 ms around the latency of the negative (between -130 and 130 ms) and preceding positive peak (-180 ms to negative peak) found in the averages. We subsequently subtracted the single-trial value of the preceding positive peak from the value of the negative peak to obtain single-trial peak-to-peak measures (same procedure as in study 2). We used an LME analysis to determine the effect of the fixed effect factor Event Type on single-trial peak-to-peak ERN amplitudes, and additionally checked effects of Empathy and Expertise with model comparisons.

In our first post-hoc analysis, we determined the effect of correction (corrected vs. uncorrected errors) on IKI, volume, and ERN amplitudes with respective LME models, and additionally checked for potential effects of Empathy and Expertise. In the second post-hoc analysis, we calculated the effect of Event Type Frequency, Difficulty and Security on IKIs and ERN amplitudes and determined, with model comparisons, if any of these variables predicted IKIs or ERN amplitudes as well as the factor Event Type.

*Results and Discussion*

We found significantly longer IKIs after large than small errors, but no difference between small errors and correct keypress. Additionally, we found significantly lower volumes for small errors than correct keypresses, but no difference between small and large errors. While pianists might be good at suppressing post-error-slowing after small errors (Jentzsch et al., 2014), large errors might be more diverting and reorienting back to the task might be more time-consuming than for small errors (Notebaert et al., 2009; Núñez Castellar et al., 2010). Respectively, corrective movements might start earlier for small than large errors.

We observed significantly higher ERN amplitudes for large errors compared to small errors and small errors compared to correct keypresses. We thus assume that error severity is represented in ERN amplitudes and thus presumably in the action monitoring system.

There was no further effect of Expertise and Empathy on any of the dependent measures. For empathy, this is in accordance with previous results concerning the monitoring of actively performed actions (Amiruddin et al., 2017). As all our participants had a musical experience of more than 500 hours and were additionally well-acquainted with the musical stimuli, we assume a potential expertise ceiling effect.

In a post-hoc analysis, corrected errors resulted in shorter IKIs, smaller volumes and higher ERN amplitudes compared to uncorrected errors (in accordance with Paas et al., 2021). We suspect that the corrective movement starts even earlier and is stronger (with even faster IKIs, possibly to make up for correction) in corrected than in uncorrected small errors, and suspect higher subjective importance of corrected compared to uncorrected errors. The effect cannot be attributed to error awareness, because participants had successfully adapted their hand movements also after included uncorrected errors. Regarding variables that might have affected expectancies, Event Type Frequency and Difficulty predicted IKIs, and Event Type Frequency additionally modulated ERN amplitudes, which might suggest some effect of expectancy, although these variables are confounded with Event Type. Both Event Type Frequency and Difficulty were significantly worse predictors of IKIs and ERN amplitudes than Event Type. Consequently, the investigated variables, and therefore presumably expectancy, cannot explain all effects of error severity, although we cannot rule out that expectancy may have some influence on own action monitoring.

***Experiment 2***

*Introduction*

A similar ERP component as for own action monitoring has been observed for observed action monitoring, which we call oMN (Bates et al., 2005; Miltner et al., 2004; van Schie et al., 2004). For observed action monitoring, previous studies (study 1, study 2) suggest that only expectancies, but not action valence affect observed action monitoring. As also no behavioral adaptation is needed in observed action monitoring, error severity might affect observed action processing less than it does own action processing. To investigate error severity effects on observed action monitoring, we recorded EEG in pianists while they watched videos of others playing (recorded in Experiment 1). In the material used for the observation study, large and small errors were roughly equally frequent, while correct keypresses were much more frequent. We therefore expected a significant oMN amplitude difference between observed correct and erroneous keypresses, but no difference between small and large errors.

For observed actions, we would expect a more pronounced effect of trait empathy than for own action monitoring due to the social component of observation, but findings are mixed (Bellebaum et al., 2020; Brazil et al., 2011; Clawson et al., 2014; Fukushima & Hiraki, 2009; Lockwood et al., 2015). Possibly, empathy shapes expectancy formation which then shapes observed action processing (see studies 1 & 2). Expertise is not always necessary for action observation (Desmet et al., 2014), but influences observed action monitoring in some ways (Panasiti et al., 2016), possibly also due to a facilitation of expectancy formation (Li & Feng, 2020; Özkan et al., 2019; Zhao et al., 2021). We therefore expected an effect of empathy, and possibly of expertise, on the oMN.

An effect of error correction on observed action monitoring is not probable, because observer participants cannot know whether an error will be corrected. However, Paas et al. (2021) found an effect of error correction on observer feedback processing. In the present study, we thus aimed to investigate this effect also for observed action monitoring. To account for factors possibly modulating expectancy, we again measured Event Type Frequency and Difficulty as well as the Perceived Expertise (how good observers rated the players in the videos) to investigate their influence in comparison to the effect of error severity. In a final post-hoc analysis, we compared the pattern of the Event Type effect on ERN/oMN amplitudes between experiment 1 and 2.

*Method*

We investigated data from 23 pianists. Participants watched videos recorded from Experiment 1 of one of the six pieces of Experiment 1 (9 videos in total, shown either 6 [8 videos] or 12 [1 video] times). Participants were required to memorize the piece in the two weeks before testing. Before the experiment, they played the piece without score notation and on a muted keyboard, had the active EEG electrodes attached (same placement as in Experiment 1) and filled out self-report measures as in Experiment 1. The experiment itself consisted of 60 sequences: in each sequence, participants saw a video (without sound), and then answered two questions, namely how many errors the player had made and how experienced they believed the player to be. After the experiment, they played the piece themselves again.

To serve as dependent behavior variable, we calculated the pre- and post-experiment performance as correct notes in percent of all notes, using a dynamic score matcher algorithm (Large, 1993; Palmer & van de Sande, 1993; Rankin et al., 2009). As dependent behavioral variable and later also as a possible predictor, we determined the Recognized Error Margin (the absolute difference between counted and actual errors). Participants were excluded if they differed more than 1.645 SD (corresponding to a percent rank < 5) from other participants in their Recognized Error Margin. For another dependent behavioral variable and as predictor in a post-hoc analysis, we recorded the Perceived Expertise as stated by participants after each trial, and took the Objective Expertise of the Player as possible modulator of Perceived Expertise into account (as stated by the person depicted in the video, indicated by self-report for experiment 1). To serve as independent variables, we calculated Difficulty (the number of times each note was played correctly in the 60 videos) and the Observed Event Type Frequency (the number of times each event happened for each observed person).

For the behavioral analyses, we determined the effect of Measurement Time (pre or post) on performance (as percent correct) using an LME model that allowed for random intercepts and slopes by Measurement Time and participant. We additionally determined an LME model with Empathy and Expertise as fixed effect factors on the dependent variable Recognized Error Margin that allowed for random intercepts by participant and video. Finally, we checked whether the Objective Expertise of the Player predicted the Perceived Expertise and additionally checked whether Empathy, Expertise or the Number of Perceived Errors explained additional variance.

EEG data preprocessing was conducted as for Experiment 1. Peak-to-peak measures were obtained similarly to Experiment 1; negative peak latencies in the averages were determined in a time window between -100 ms and 100 ms. This is earlier than for previous studies investigating the oMN, although latencies seem to differ depending on the task (Bates et al., 2005; Miltner et al., 2004; van Schie et al., 2004). As in active sequential tasks (Maidhof et al., 2013), the observed movements might also be detected earlier (see Di Gregerio et al., 2022) when observing sequential tasks, resulting in earlier oMN latencies. The preceding positive peak was determined in a window between -150 ms and the negative peak latency. Values were extracted in the single-trial data as the mean value 10 ms before to 10 ms after the average peak latency, and the single-trial positive peak was subtracted from the single-trial negative peak to obtain single-trial peak-to-peak oMN values.

For the EEG main analysis, we determined an LME model with oMN amplitude as dependent variable and Observed Event Type as fixed effect factor. We allowed random intercepts by participant and video. With model comparisons, we checked for additional influences of Empathy, Expertise and Recognized Error Margin.

In the first post-hoc analysis, we calculated a similar LME model with Correction as fixed effect factor, also checking for additional influences of Empathy, Expertise and Recognized Error Margin. In the second post-hoc analysis, we determined the influence of Observed Event Type Frequency and Difficulty on oMN amplitudes and checked whether Observed Event Type predicted oMN amplitudes better than the two variables. Additionally, we investigated whether more variance was explained if perceived expertise of the players in the video was included into the main analysis model to predict oMN amplitude.

In a final post-hoc analysis, we created a model with Event Type and Agency (active, observed) as fixed effect factor and z-transformed ERP amplitude values (single-trial ERN for the active, single-trial oMN for the observer group) as dependent variable. Random intercepts were allowed per participant.

*Results and Discussion*

Participants performed the piece with an average accuracy of around 85%, and there was no significant difference between pre- and post-test. As performance did not improve after watching 60 repetitions of the piece, we assume that participants did not learn additionally by watching the videos. The Recognized Error Margin did not depend on Empathy or Expertise, and participants differed on average 5.6 errors from the actual error number. A small variance between participants and no influence of interindividual measures suggests that this high number probably does not result from low performance in error monitoring, but from participants using other criteria to count errors than the score matcher algorithm. To account for potential interindividual differences in error recognition, we included Recognized Error Margin as a potential factor in the main analysis. However, the variable did not lead to increased model fit in the main model. Perceived Expertise was not influenced by actual expertise, but by the perceived number of errors.

Concerning the oMN, we found higher amplitudes for small errors compared to correct keypresses, but no difference between small and large errors. This suggests that there was no or at least a smaller difference between small and large errors than for own errors. The missing effect might be attributed to the fact that the type of error was not relevant for the observed action monitoring task.

For the EEG main analysis, we again found no influence of empathy or expertise. Expectancy formation concerning observed performance accuracy might have been easy in the current task. The high expertise of observers (at least 780 h), together with the fact that they knew the piece by heart, might have further facilitated expectancy formation (Li & Feng, 2020; Özkan et al., 2019; Zhao et al., 2021), leading to very low variance in expectancies and, respectively, no empathy or expertise effects.

We found no difference between the processing of observed corrected and uncorrected errors. The oMN in our study occurred around the keypress itself and the mean interval between the error and the corrective keypress was almost three times as long as in the study by Paas et al. (2021), so an effect of the next keypress on the oMN (correction perceived as two subsequent errors) was improbable. In contrast to our study, Paas et al.'s participants also heard the volume of the note. In Experiment 1, we found that volume differed significantly for corrected compared to uncorrected actions, and such an effect might have been perceived as surprising or as an indicator of higher error significance by the observers in Paas et al.'s study.

We found significant effects of Observed Event Type Frequency and of Difficulty on oMN amplitudes, and Observed Event Type Frequency was even as good a predictor as Observed Event Type. Observed Event Type Frequency and Observed Event Type were confounded and we cannot attribute the effect definitely to either of the predictors, but the result matches the PRO model's claim (Alexander & Brown, 2011) that valence effects in action monitoring could be explained by the low frequency of errors. Finally, we found no effect of perceived expertise on error processing. As this variable was also not predicted by the actual expertise, participants might have been insecure about making expertise assumptions and the weak assumptions did not affect expectancies and/or oMN amplitudes.

In the third post-hoc analysis, we found a significant interaction between Agency and Event Type on z-transformed ERP amplitude values, confirming that the pattern of the effect of Event Type differed significantly between active and observer conditions. This difference was driven by the effect of large errors: only for this event, we found a significant difference between active and observer groups, with significantly higher z-score ERP values for the active group. This confirms that an effect of error severity is present only for own, but not for observed action monitoring.

***Conclusion***

We found an effect of error severity on own, but not observed action monitoring. Especially regarding effects of expectancy (which can explain results for observed, but not own actions), we propose that a need-to-adapt signal is sent by the action monitoring system when either predictions or actions need to be adapted. In the present study, action adaptation was needed only for Experiment 1, so it is perceivable that results in Experiment 2 are based (only) on prediction adaptation. Further research should test different levels of errors, and the effect of consequences (if adaptation is needed or not) on action monitoring.

## General Discussion

The three studies included in this dissertation addressed in detail several modulators of own and observed action monitoring, namely expectancy, empathy, and error severity. In study 1, we used a false-belief paradigm with an additional difficulty modulation to disentangle vicarious error processing and expectancy processing in observation. In study 2, we used a similar false-belief paradigm to disentangle effects of expectancy and empathy on observed action monitoring. In study 3, we investigated the effects of error severity on own and observed action monitoring by means of a piano playing paradigm while additionally investigating modulations by empathy, expertise and expectancy.

In study 1, we modified observer participants' expectancies using false-belief and task difficulty. Effects of empathy and expertise interacted, with effects in an early time window of both modulators in the conditions in which false-belief was relevant and effects of just expertise in the conditions where it was not. A similar pattern emerged for participants' behavioral expectancies.

In study 2, we used a similar paradigm as in study 1, which included only a false-belief-, not a difficulty manipulation. We found that only for participants with low empathy, expectancies grew weaker for the false-belief condition during the experiment, suggesting a fatigue effect. We used regression lines to estimate single-trial expectancies for each participant and belief condition and found that the expectancy values predicted amplitudes of the oMN. This effect was not further modulated by empathy. Based on study 1 and 2, we assumed that empathy shaped expectancies which then shaped observed action monitoring.

In study 3, we found that own, but not observed action processing was influenced by the degree of deviation of an error from the correct movement, i.e., larger errors elicited higher ERN amplitudes than small errors. Effects of expectancy seemed to be stronger in observed than own action processing, although valence and modulators of expectancies were confounded.

In the following paragraphs, I will discuss possible modulators of neural action monitoring as investigated in the three studies, with a focus on expectancy, empathy and error severity.

**Expectancy**

After the formulation of the PRO-model (Alexander & Brown, 2011), evidence accumulated confirming a strong effect of predictions and expectancies on own and observed action monitoring (Bellebaum et al., 2020; Chase et al., 2011; Ferdinand et al., 2012; Gawlowska et al., 2018; Jessup et al., 2010; Kobza et al., 2011; Kobza & Bellebaum, 2013; Notebaert et al., 2009; Núñez Castellar et al., 2010; Oliveira et al., 2007; Schiffer et al., 2014; Wang et al., 2015; Wessel et al., 2012). Studies 1 and 2 confirm this effect on an electrophysiological measure of observed action monitoring, i.e., the oMN, and further investigate how predictions are formed. In study 1 and 2, we found no main effect of valence on observed action monitoring when expectancies were accounted for. Expectancy was not directly modulated in study 3 and variables potentially influencing expectancy (event type frequency and difficulty) could not be disentangled from valence effects because events differed significantly in their frequency and difficulty. Main effects of frequency on ERN amplitudes and post-event reaction times (Experiment 1) as well as oMN amplitudes (Experiment 2), and main effects of difficulty on post-event reaction times (Experiment 1) and oMN amplitudes (Experiment 2), nevertheless suggest that an expectancy modulation cannot be ruled out. Accordingly, at least studies 1 and 2 clearly confirm the role of expectancies on observed action processing: they showed that observed action processing is independent from vicarious error processing (see study 1) and that another modulating factor on observed action processing, namely empathy (study 2), might even influence expectancy formation instead of observed action processing directly.

As mentioned above, we found no effect of valence on oMN amplitudes in studies 1 or 2 (in accordance with Bellebaum et al., 2020; Ferdinand et al., 2012; Oliveira et al., 2007; Schiffer et al., 2014). However, in a number of studies, even if a significant effect of expectancies emerged, valence did seem to play a role (Gawlowska et al., 2018; Jessup et al., 2010; Núñez Castellar et al., 2010; Pezzetta et al., 2018; Wang et al., 2015) – an observation we also made in study 3. Some of the inconclusive results of previous studies can be accounted for by an overoptimistic bias of participants concerning their and others' performance (Oliveira et al., 2007) which we also observed in study 1. However, our results from study 3, as well as corroborating results on the effects of motivation and error type (Ganushchak & Schiller, 2008; Hajcak et al., 2005; Maier & Steinhauser, 2016), suggest that expectancies are not the only factor shaping action monitoring, and, at least for own action monitoring, an action valence effect on action monitoring cannot be ruled out. For observed action monitoring, valence might play a role only if the other persons' actions have any implications for the observed person, for example if their responses lead to winning or losing money (Marco-Pallarés et al., 2010) or if they are watching goals from their supported or an opposing team (Newman-Norlund et al., 2009). In all three studies comprised in this dissertation, the observed person's actions had no consequences for the observer participants. Instead, the task was to observe the person as thoroughly as possible to be able to answer prompts. In studies 1 and 2, observers had to for expectancies about the observed responses, in study 3, observers had to monitor the observed person's performance and expertise (study 3). Our observer participants' intention, therefore, was not for the observed person to answer in a certain way, but to be able to observe, and thus predict, their movement as best as possible. In this sense, 'valence' and 'expectancy' might even refer to the same construct for our participants. If they were able to predict the observed person's answer, the observer acted correctly, and if not, they made a (prediction) error.

As a limitation regarding the findings on expectancy, in studies 1 and 2, we measured expectancy directly with prompt questions interspersed throughout the experiment. However, we did not acquire expectancy values on a single-trial basis (because this would have possibly changed participants' attention). This only allowed us to approximate actual expectancy in single-trial values in study 2 (and averaged values in study 1), although these approximations already sufficed to find a respective effect. In study 3, we did not measure expectancies directly, but via event type frequency, note difficulty and security. While some of these (note difficulty and security) were acquired on a single-note or single-trial level, we can only infer potential expectancy effects from these variables, and they were additionally confounded with valence. In the future, it might be interesting to directly ask for participants' expectations after each respective trial.

Having said that, it is questionable whether expectancies can be fully accounted for by explicit measures of expectancy. In study 1, for example, we found enhanced amplitudes for correct actions in the high-difficulty true-belief condition and enhanced amplitudes for erroneous actions in the high-difficulty false-belief condition, which was not in accordance with behaviorally measured expectancies. As a block design for the false-belief condition was used, relative comparisons of the difficulty conditions (answers should be more likely to be correct in the high-difficulty trials compared to low-difficulty trials in the false-belief blocks, and vice versa in the true-belief blocks) could explain the result pattern. These expectancies could have been implicit, as they were not reflected in the explicit expectancy measures. In future studies, measures of explicit expectancy might be complemented with behavioral measures (reaction times and accuracy data) that could account for (implicit and explicit) expectancies.

In summary, studies 1 and 2 clearly show that expectancy, not observed action valence, is coded in the action monitoring system when observing others (at least when the valence of the observed action has no implications for the participants). In study 3, although event type frequency and valence cannot be disentangled, the results for the oMN could be explained by event type frequency as well as by valence. This does not directly support an expectancy effect, but still supports the PRO model's (Alexander & Brown, 2011) statement that most valence effects can be explained by error frequency. For own action monitoring – where the valence of an action almost always has implications for the participant – the (sole) influence of expectancy is more debatable. Error significance modulated ERN amplitudes in study 3 even though the error types did not differ in their frequency, and variables that might influence expectancy formation could not explain this effect. Expectancies in observed action monitoring might be shaped by multiple factors, for example by the comparison between conditions (see study 1). A specifically important modulator of expectancies, as observed in studies 1 and 2, seems to be trait empathy.

**Empathy**

Effects of trait empathy on own action monitoring have been considered improbable (Amiruddin et al., 2017), although the role of the ACC in both action monitoring (see Ridderinkhof, Ullsperger, et al., 2004; Taylor et al., 2007) and empathy (Bernhardt & Singer, 2012; Lamm et al., 2010; Lamm et al., 2011; Singer et al., 2004) suggests a possible connection between the two. In accordance, we found that empathy did not influence own action monitoring in study 3. Presumably, effects of action monitoring are either modulated by different subareas of the ACC, or they are similar, but unrelated, processes.

In contrast, effects of empathy on observed action monitoring are more likely, because of the similarity between brain responses for own and observed emotional states or actions (Bates et al., 2005; Bufalari et al., 2007; Lamm et al., 2011; Miltner et al., 2004; Singer et al., 2004; van Schie et al., 2004). A number of studies corroborate this assumption (Bellebaum et al., 2020; Brazil et al., 2011; Fukushima & Hiraki, 2009; Lockwood et al., 2015; Newman-Norlund et al., 2009; Shane et al., 2009). However, findings are inconclusive. First, there are also observation studies that found no effect of empathy at all (Clawson et al., 2014; study 3). Second, the aforementioned studies found correlations with different sub-constructs of empathy (measured with different subscales of empathy questionnaires). Of these studies, only Fukushima and Hiraki (2009) and Bellebaum et al. (2020) investigated general empathy and only Bellebaum et al. (2020) found an effect of general empathy.

In studies 1 and 2, we therefore aimed to further investigate a potential effect of empathy on observed action monitoring. For false-belief tasks, general empathy seems to be important to form predictions (Bellebaum et al., 2020; Ferguson et al., 2015), and we indeed found that general empathy, as measured with the Cambridge Behavior Score (Baron-Cohen & Wheelwright, 2004), influenced behavioral measures of expectancy as well as observed action monitoring in both studies. Subsequently, I will discuss the effect of general empathy on expectancy formation, and then discuss the consequent influences of empathy on observed action monitoring.

In both studies, higher trait empathy led to improved expectancy formation. In study 1, this emerged as stronger expectancies for high empathy participants compared to low empathy participants only in the condition in which false-belief reasoning was necessary to form expectations. In study 2, the strength of expectancies decreased across the experiment only for low empathic participants in the false-belief condition. In study 1, presumably because the task was more complex, the empathy effect emerged across all trials, while in study 2, in a less complex task, the effect emerged as a fatigue effect in later trials. No effect of empathy on expectancies was found in a previous study with the same paradigm as in study 2 (Bellebaum et al., 2020), but effects were probably masked both by a homogenous sample, resulting in an expectancy ceiling effect, and the fact that variations across the experiment were not taken into account. The role of empathy on performance in a false-belief task is well established (Birch & Bloom, 2007; Wellman et al., 2001), and tasks with a social component (Brown & Brüne, 2012) might require empathy for expectancy formation, while others do not. Accordingly, empathy modulated expectancy formation in conditions in which the social cognitive load was especially high in studies 1 and 2. We interpreted the result patterns in the sense of a social working memory (Meyer et al., 2012; Meyer & Lieberman, 2012) of which more or less resources are needed to perform a task (see Apperly et al., 2008; Epley et al., 2004). For people with low trait empathy, the social working memory might be impaired, resulting in overall decreased performance (study 1) or stronger effects of fatigue on performance (study 2).

Based on the similar pattern between behavioral expectancy data and electrophysiological results concerning empathy effects in study 1, we suspected an indirect effect of empathy on neural correlates of observed action monitoring via expectancy formation. This was confirmed by a post-hoc analysis in which we found that behavioral measures of expectancy influenced oMN amplitudes, but no further effect of empathy. In study 2, as well, expectancy values influenced amplitudes of the oMN. Extending study 1, this was shown on a single-trial level, and again, empathy did not explain any additional variance. We suspect, based on the results from studies 1 and 2, that empathy has an indirect effect on observed action monitoring via its effect on expectancy formation. Depending on how much – or which aspect of – empathy is needed to form expectancies, empathy can thus play a larger or smaller role in observed action monitoring.

In accordance with this assumption, we found no effect of general empathy on observed action processing in study 3, when participants observed other persons play a piano piece the observers knew by heart. Both their high experience with the piece and the first-person perspective of the used videos (Angelini et al., 2018; Drew et al., 2015) probably made expectancy formation effortless. This was confirmed by the fact that empathy did not influence participants' ability to count the players' errors in the videos. Also, expectancies, as opposed to studies 1 and 2, were not manipulated systematically, especially not by a social-cognitive demanding manipulation such as false-belief. This again might have led to easier and possibly more varied expectancy effects. Participants' expectancies might have been modulated by the perceived expertise, number of errors of a specific observed person, and the knowledge about difficult passages in the piece. As no focus was put on these expectancies (as opposed to a direct instruction about what observer participants should expect in studies 1 and 2), many underlying factors could have played a part in expectancy formation (e.g., the sympathy felt towards specific players, the speed in which the piece was played). As a result, the empathy and expertise effects on expectancy formation, and thus on observed action monitoring, might have been respectively diminished.

As a limitation regarding the effects of empathy in the three studies, we used videos or simulations for the observation condition in all three studies. In studies 1 and 2, participants were told that they saw videos of a real person, but trials were simulated. In study 3, participants did observe a video of a real person, but still, no real-life observation happened and empathic processing might have been diminished by the simulation (studies 1 and 2) or by the first-person perspective in study 3. As effects of empathy are reduced when watching computer actions as opposed to other person's actions (Fukushima & Hiraki, 2009), and empathy recruitment seems to depend on how social a task is perceived (Brown & Brüne, 2012), presenting simulations and videos on a monitor might compromise empathic processing. Thus, effects of empathy might be investigated more efficiently in a more real-life setting where participants observe other persons in the same room.

Another problem with regard to empathy is that both the conceptualization and measurement of the concept are highly debated in the literature. No consensus exists on a definition of empathy or the concepts and components it contains. It has even been suggested to avoid the use of the unspecific term 'empathy' altogether in favor of specific components that are believed to be a part of empathy, such as perspective taking (Hall & Schwartz, 2019). We used a self-report measure of empathy, using the Cambridge Behavior Scale, to investigate effects of empathy in all three studies (Baron-Cohen & Wheelwright, 2004). This measure has been connected to processing in false-belief tasks (Bellebaum et al., 2020; Ferguson et al., 2015), making it the preferred measurement for our paradigms in studies 1 and 2. The questionnaire controls for response bias by using positively and negatively scaled items and, in some way, for social desirability bias by introducing 20 distractor items to avoid a focus on empathy. However, the authors themselves note that the self-measure only depicts the individual's belief – or desired belief – about themselves. In addition, the questionnaire provides a general empathy measure and therefore is based on a possibly unjustified definition of empathy and its underlying concepts. Previous studies used a wide range of self-report measures for empathy that do not necessarily correlate with each other (Hall & Schwartz, 2019). In a meta-analysis, Wright et al. (2021) found that the magnitude of impairment of affective empathy across neurological diseases differed relative to which measurement was used (self-reports, other-reports and performance). Only for strong overall empathy effects in specific neurological diseases, all measurements led to the same result pattern. This suggests that for some – possibly less pronounced – empathy effects, different types of measurement might lead to conflicting results. In the future, a combined empathy measure (including self-report, other-report, performance and physiological measures) should be created and validated to be used in further studies. Such a measure might be used to gain an encompassing picture on empathic effects. Also, the focus in future research could lie on more specific concepts that are presumed to be a part of the general concept of empathy.

In summary, empathy might affect expectancy formation in observation, thereby having an indirect influence on neural correlates of observed action monitoring. The amount and characteristic of empathy needed for expectancy formation seems to depend on the social-cognitive demands of the respective observation. Further research should expand on this assumption by using in-person observations and improved empathy measurements.

**Error Severity**

As established above, the PRO model cannot explain all findings on action monitoring, especially the influence of motivation for good performance (Ganushchak & Schiller, 2008; Hajcak et al., 2005). These findings suggest an effect of subjective error significance on action monitoring. They are corroborated by studies finding altered amplitude sizes for different error types (Bernstein et al., 1995; Falkenstein et al., 2000; Maier et al., 2012; Maier & Steinhauser, 2016; Murata & Katayama, 2005; Paas et al., 2021), even if frequencies of error types are accounted for (Maier & Steinhauser, 2016). We therefore suspected that in addition to expectancies, the deviation from the correct movement would influence action monitoring in a continuous way. We indeed found that large errors elicited larger ERN amplitudes than small errors in a piano-playing paradigm in study 3. Also, small errors were played significantly more quietly than large errors and correct keypresses, but did not lead to post-error-slowing. Contrastingly, large errors did not lead to volume change, but to post-error slowing, which suggests that the processing speed of small and large errors differed (small errors led to earlier corrective movements). We additionally replicated findings by Paas et al. (2021) showing that ERN amplitudes for later corrected errors were increased compared to uncorrected errors. Although the specific mechanisms behind this effect are not yet investigated, we can rule out an effect of error recognition. This is because uncorrected errors were only included if they were followed by a correct note, meaning that the subsequent hand movement after the error had to be (intentionally) adapted. Participants were specifically instructed to not correct their errors and probably suppressed this reaction in uncorrected errors, but failed to do so in corrected errors. In a study investigating behavioral effects of errors in skilled typists, Crump and Logan (2013) observed that error correction happened rapidly and even led to post-error speeding, while post-error slowing occurred after errors when correction was disallowed. This suggests that for experts of sequential tasks, error correction avoidance might require additional cognitive resources. We therefore suspect that corrected errors led to an attentional focus on the error and, respectively, less attention on the task (as proposed by Notebaert et al., 2009), which then led to a failure to adhere the no-correction instruction. Presumably, this attention deviation happened specifically for corrected errors because they were perceived as specifically significant. This suggests that ERN amplitudes are also influenced by the subjective importance of errors.

Please note that the processing of error severity was crucial for the task used in study 3, because piano playing is a sequential task, meaning that corrective actions of some type have to be conducted to continue fluent play, and these actions have to be more extensive for large than for small errors. We only included events where the previous and following note were played correctly to avoid overlapping processes, so the corrective actions after included errors were always successful. Event type (correct, small and large error) predicted both post-event reaction times and ERN amplitudes better than variables related to expectancy (event type frequency, difficulty and security). The effect of expectancies cannot be ruled out and has not been manipulated in the study. However, there are also valence effects, specifically of the effects of error severity, that presumably cannot be explained by expectancy, at least when error severity is relevant for task performance.

For observed action monitoring, amplitudes did not differ significantly between large and small errors or between corrected and uncorrected errors, and effects could be explained by event type frequency as well as by event type. A post-hoc analysis revealed that the effect of event type on the oMN was significantly different than that on the ERN: while amplitudes (z-scored, to make the two measures comparable) were similar for correct events and small errors, for large errors, amplitudes for own action monitoring were significantly larger than amplitudes for observed action monitoring. In the observation task, error severity was not relevant for task performance because observer participants, firstly, did not perform any corrective actions and, secondly, were instructed to count all errors regardless of their severity. Based on our results, we suspect that the need for adaptation, and the size of the respective adaptation, is also (besides expectancies) coded in action monitoring.

In conclusion, for own action monitoring, error severity, and possibly also subjective error importance, influences neural mechanisms of action monitoring as measured in ERN amplitudes. This effect cannot be explained by event type frequency or other measures that might influence expectancies. For observed action monitoring, differently severe errors do not lead to significantly different oMN amplitudes, and an influence of expectancy cannot be ruled out. These results have to be considered on the basis that for own action monitoring, valence and specifically error severity were important information to adapt hand movements, while the significance of errors had no consequence for observers in observed action monitoring.

**Need-to-Adapt Signal**

Action monitoring as a process serves to control our actions and their outcomes as much as possible, and if they are not as they should be, adapt them (as in reinforcement learning, see Holroyd & Coles, 2002). But, the phrase "as they should be" already raises a problem which was investigated in action monitoring research in recent years: does that mean "as they are expected" or "as they are desired"? For successful control, clear expectancies as to actions and their outcomes are necessary, and the brain should strive to adapt predictions whenever they are not met. As the PRO model states, mPFC activity should reflect the unexpectedness of actions or outcomes rather than their valence, and we found corresponding results for observed actions in studies 1, 2, and to some degree, study 3.

However, action monitoring also seems to be dependent on subjective goals: increased neural error monitoring responses have been observed for participants with high sensitivity for errors (with identical or even increased expectancy of errors), as in anxiety patients (Carrasco, Harbin, et al., 2013; Endrass et al., 2014) and highly perfectionist participants (Barke et al., 2017; Stahl et al., 2015). Neural responses also increase when the motivation for good performance is high (Ganushchak & Schiller, 2008; Hajcak et al., 2005). Adapting expectations might not always be enough to exhibit maximum control; the brain needs to take subjective goals and deviations from them into account. If a pianist wants to play a piece perfectly, it does not suffice that they expect the error (that is made at the same specific note every time they play the piece), they need to adapt their actions to avoid the error in the future – which is what we showed in study 3. We proposed in study 3 that the mPFC, as a region involved in action monitoring and reinforcement learning (Holroyd & Coles, 2002) sends a general need-to-adapt signal. When predictions are not met, the brain needs to adapt to the new environment (in accordance with the PRO model). Additionally, when behavior deviates largely from the desired action, adaptation might also be necessary, depending on the subjective goal. Brain activity in action monitoring might thus code the need to adapt, not just for prediction models, but also for action models. In study 1, we found that the magnitude of what we believe to be an adapt-signal was dependent on the size of the prediction error, i.e., the more the observed actions differed from what was expected, the larger the monitoring response, because the more adaptation of the prediction model was needed. This was also observed for feedback processing (Ferdinand et al., 2012). The adaption signal thus should represent the need to adapt not only as a dichotomic measure, but indicates the magnitude of adaption needed to reach the desired action or prediction model. Support for this assumption stems from studies indicating a correlation between FRN amplitude and prediction error on a single-trial basis (Burnside et al., 2019; Fischer & Ullsperger, 2013; Krigolson et al., 2014), although the effects found in these studies cannot be interpreted independently of valence, as the authors calculated a signed prediction error.

In the observational tasks used in study 1 to 3, we found a signal that could be accounted for by prediction errors only. In all paradigms, errors had no consequences for the observer participants, and observer participants were not able to influence the observed actions, so action models had little to no significance for them. On the contrary, their specific task was to closely observe and draw conclusions regarding the observed persons' next move (studies 1 and 2) or their performance (study 3), so it is conceivable that prediction model adaptations were of particular relevance for them. The observed responses can even be seen as the feedback to the observers' predictions and reflect how good participants are at the task and how much they need to adapt. Observed action monitoring might thus be only a specific task for a general action monitoring system that works to improve task performance. In accordance, if observed outcomes are relevant for the observer, as in studies setting a cooperative or competitive environment, neural responses to others' errors or negative feedback were shaped by the meaning of the outcome for the observers (Koban et al., 2010; Marco-Pallarés et al., 2010; Newman-Norlund et al., 2009) and presumably not just by prediction errors. In the active playing condition of study 3, both predictions and action models were of relevance for the participants to continue fluent play, and accordingly, the need-to-adapt signal reflected action model deviations (presumably in addition to prediction model deviations, although we cannot conclude this from our results with any certainty).

In summary, we suggest that both expectancies *and* action valence might influence a general action monitoring system (for both own and observed action monitoring) depending on what is relevant for successful task completion in a continuous, non-dichotomist fashion. This assumption should be tested empirically in future studies by investigating different levels of valence (partial errors, small errors, large errors, very large errors) as well as different levels of expectancy (high, medium, and low error expectancies) in both own and observed action monitoring. Especially for the valence effect, this is not easy, because new paradigms have to be developed to test different levels of action valence in an ecologically valid way. As the piano keyboard proved an adequate instrument in study 3 because of the normed key width, we suggest to use a piano-like device in (non-musical) future studies, for example in a stimulus-reaction-task where different stimuli are mapped to different keys. Participants would have to move their hand from a fixed point in front of the keyboard to the respective key as fast as possible, which might induce smaller and larger errors. With ever-improving measurements of 3-D movement kinematics (Marshall et al., 2022), these could be used for analyzing specific sports-related movements and their deviation from the correct action. As an example, dance movements requiring specific foot placements could be practiced by participants, and then performed while 3-D movement kinematics are reported. From these data, deviations between actual and correct foot placement could be calculated.

Participants might differ in how severe they perceive an error, so individual error attributions should be taken into account. This could be done by asking participants after each trial if they believed they performed correctly, and if not, how far they believed their action deviated from the correct movement.

Videos, simulations or real-life observations of the same tasks could be used for observed action monitoring. Expectancy can be manipulated stepwise via task difficulty or via emphasizing speed on specific tasks in own actions, and via task difficulty, speed variances or instructions in observed actions. In an adaptation-paradigm, participants' performance could be manipulated so that the task difficulty (e.g., by visibility of stimuli, see Notebaert et al., 2009) is increased if they are below the desired error frequency, and decreased if they are above. Regarding observed action, we were able to manipulate expectancies in study 1 by adapting task difficulty, making observer participants belief that the player would perform less well in some conditions. A similar effect might be achieved by manipulating expectancies with instructions concerning players' ability for this task ("this person is especially good at this task"/"is less good").

For own and observed action monitoring paradigms, the relevance of prediction and action models could be manipulated in future studies. Previous studies showed that increasing reward led to higher ERN amplitudes (Ganushchak & Schiller, 2008; Hajcak et al., 2005). In further studies, the interaction between significance of a specific error for the outcome (monetary gain/punishment) and the error severity might be investigated. This could be done by comparing differently sized errors that are either punished depending on the deviance of the actual from the desired action, depending on their absolute valence (correct or wrong) or not at all. For own action monitoring, it might also be interesting to compare sequential and non-sequential tasks, because immediate corrective movements are only needed in sequential tasks. For observed action monitoring, the possibility for corrective actions might also be manipulated in a joint-action task (e.g., Loehr et al., 2013; Paas et al., 2021) – if correction is possible, again, more emphasis should be on action model adaptation. In conclusion, modulations of both the importance and magnitude of prediction and action errors should be investigated in future studies to explore the implications of the need-to-adapt theory.

Finally, while we established that empathy influences the size of the need-to-adapt signal for prediction models in observed action monitoring by manipulating the strength of predictions, we can only presume its role for adapting action models. Possibly, empathy, at least in specific contexts, could influence the motivation to answer correctly or see another person answer correctly (as suggested by cooperative/competitive context studies, Marco-Pallarés et al., 2010; Newman-Norlund et al., 2009). We suggest accounting for empathy effects in future studies investigating the assumptions of the need-to-adapt signal, especially to test its effects on the processing of action valence, not just expectancies.

In short, the results from studies 1-3, but specifically from study 3, leave us to assume that the action monitoring system generates a need-to-adapt signal whenever an action or prediction adaptation is needed for successful task completion. This signal includes the size of the adaptation needed, i.e., is larger for large required adaptations compared to small adaptations. Depending on the task, the signal might code either prediction adaptations, action adaptations or both.

**Expertise**

Besides findings on three important modulators of action monitoring, namely expectancy, empathy and error severity, which were the main focus of this dissertation, the studies 1 to 3 included some additional findings and methods which will be discussed in the subsequent paragraphs. First, we also investigated expertise as a possible modulator of action monitoring in study 3 – more on a side note, because we used an expert sample -, and I will shortly discuss the results.
In study 3, we found no effects of expertise on own and observed action monitoring. Expertise has been shown to modulate own (Harris et al., 2014; Jentzsch et al., 2014; Rachaveti et al., 2020) and observed action monitoring (Candidi et al., 2014; Panasiti et al., 2016). We suspect that expertise has a similar effect on action monitoring as empathy. While expertise might not always be necessary to form predictions and thus might not be necessary for action monitoring (Desmet et al., 2014), it might, in some aspects, be beneficial to form expectancies about own and others' actions (Özkan et al., 2019; Zhao et al., 2021). However, in study 3, we investigated an expert sample, and while participants differed in their expertise, all had more than 500 hours of experience with piano

playing and were also experts for the concrete material used in the study as they practiced the pieces before the experiment. This could have resulted, as mentioned earlier, in effortless expectancy formation and in an expertise ceiling effect. Additionally, no comparison with novices was possible.

As another explanation for the missing effect, we measured expertise as the total number of hours participants spent with the instrument, but other aspects, such as genetic predisposal and practice techniques have to be considered when investigating expertise (Bonneville-Roussy & Bouffard, 2014; Meinz & Hambrick, 2010; Williamon & Valentine, 2000). Wallentin et al. (2010) suggest to use behavioral musical performance as a measure of expertise and developed the Musical Ear Test. In future studies investigating (piano playing) expertise, behavioral performance tests might be developed for a more direct measure of expertise.

On another note, effects of expertise on action monitoring might be different depending on the type of expertise. For the task on which expertise has been acquired, expertise should lead to less errors and improved action-outcome predictions in own and observed errors. However, it is possible that to achieve expertise in some areas, action monitoring itself is important. As an example, in musical training, errors should be avoided, but when they happen, students also need to react quickly and effectively to them (Kruse-Weber & Parncutt, 2014). Expertise, at least in some areas, might thus equal increased experience in error management. In accordance, Jentzsch et al. (2014) found that musical training led to improved action monitoring. Future studies could compare the effects on action monitoring of different areas of expertise in which either error behavior is specifically practiced (e.g., music and most sports) or not (e.g., running, for which success is shaped less by movement perfection and more by fitness and constant practice).

To conclude, we did not find an effect of expertise on own or observed action monitoring in study 3, which might be due to an expertise ceiling effect in our study. Nevertheless, expertise as a modulator of action monitoring should be further investigated, especially for areas in which expertise might include a specific expertise in error monitoring and management. For future studies, expertise should be measured with a behavioral performance test to obtain objective values.

**Latencies of the oMN**

Moving away from (observed) action monitoring modulators to the component reflecting the monitoring process, we made some interesting observations regarding the latency of the oMN. In most electrophysiological studies on observed action monitoring, the respective ERP component that we call oMN peaks between 100 to 250 ms after the onset of the observed response (Bellebaum et al., 2020; Carp et al., 2009; de Bruijn & von Rhein, 2012; Kobza & Bellebaum, 2013; van Schie et al., 2004), although earlier peaks have been observed for go/no-go tasks (Bates et al., 2005; Koban et al., 2010). In the three studies included in this dissertation, we found different latencies of the oMN: In study 1, we found effects in an area of 100 to 250 ms after the observed action; in study 2, we found effects in an area of 250 to 420 ms; and in study 3, we found effects in an area between 100 ms pre- and 100 ms post-event. These discrepancies suggest that oMN latencies are highly dependent on the task, or more specifically, on the time point at which observer participants receive and process relevant information. In study 1, we used a block design for the false-belief condition, and the second factor, difficulty, was visible at the very beginning of the trial. As a consequence, participants had already formed their expectancies and just had to compare them to the observed action at the time of the response of the observed player. This might have led to earlier processing compared to study 2, where the order was randomized and participants knew only at the end of the trial whether it was a false- or true-belief trial. In study 3, we used a sequential task. Research from own action processing suggests earlier ERN latencies for sequential tasks (Herrojo Ruiz et al., 2009; Kalfaoğlu et al., 2018; Maidhof et al., 2009; Paas et al., 2021), possibly because actions start earlier as in non-sequential tasks and errors are thus noticed earlier as well (Maidhof et al., 2013). A similar effect is conceivable for observation, because also here, participants can assume the action outcome already when the player is moving the finger to the respective key, before the actual keypress. In previous studies using observation in sequential tasks (joint-action studies using music performance, Loehr et al., 2013; Paas et al., 2021), participants observed others only by listening to their musical output. In this setup, observer participants did not have the advantage of observing early movement indications, and latencies were more comparable to non-sequential observational tasks.

The three studies of this dissertation are exemplary for the variability of the oMN latency. Presumably, the oMN is generated as soon as all information needed to interpret the action is collected, and, depending on the conditions, this can be earlier or later.

**Single-Trial Approach for Obtaining ERP Peak-to-Peak Amplitudes**

As a third additional point, I want to discuss the novel approach for obtaining single-trial ERP amplitudes that we developed for study 2 and subsequently used in study 3. Analyses from study 2 showed that single-trial results were comparable to results from averaged data (compared to Bellebaum et al., 2020). We calculated single-trial data by first finding the maximum negative and preceding positive peak in the averaged data (averaged by participant and condition), a step identical to conventional average EEG analyses (as also used in study 1). However, instead of retrieving the amplitude values at the respective latencies in the averaged data, we retrieved values at the respective latencies in each single segment. In study 2, to achieve maximum comparability with previous studies (study 1; Bellebaum et al., 2020; Kobza & Bellebaum, 2013), we extracted only the value at the peak point. In study 3, we extracted an area in a -10 to +10-ms area around the peak to account for slight latency differences also within-subject and within-condition.

The chosen single-trial approach offers two main advantages: first, it is a relatively easy and straightforward method to account for single-trial variances in EEG signals, and we could use the same statistical analysis method for single-trial analysis in study 2 and 3 as used for average analysis in study 1, namely LME models (LME models work with both data types). Second, the average from all extracted single-trial data would be the same value as used in conventional average studies, so the approach allows comparability with previous studies while still including more data points.

Nevertheless, a further progress in EEG data analysis is desirable, because even though our developed method includes much more data than previous analyses, it still only includes a fraction of all acquired EEG data. Also, some problems of average analyses remain, for example that within-subject trial-to-trial peak latency changes are not accounted for and a large variance in latencies can lead to smaller peaks and reduced effect sizes (see Luck, 2014). In a time with ever-improving technological and computational solutions, more and more possibilities to optimize EEG analyses by including an increasing number of data points arise. LME models seem a good approach for single-trial analysis (see also Frömer et al., 2018; Spinnato et al., 2015), and advances in machine-learning also suggest a promising role for the method in future EEG analysis methods (see Stewart et al., 2014; Wirth et al., 2018).

**Conclusion**

The functional role of the action monitoring system and its modulators is not sufficiently clarified, especially for action observation. We aimed to investigate possible modulators of own (study 3) and observed (study 1, 2 and 3) action monitoring. In studies 1 and 2, we found that prediction formation in observed action monitoring is modulated by empathy, and these predictions then modulate observed action processing, with larger amplitudes of negative frontocentral ERP components for less expected observed actions. Importantly, we found no effect of observed action valence. In study 3, we found that error severity further modulates own action monitoring, and that this effect cannot be explained by participants' expectancies. Opposed to that, we found no effect of error severity in observed action monitoring and additionally found that the valence effect could also be explained by expectancies. Based on the findings of studies 1 to 3, we propose that the action monitoring system sends a need-to-adapt signal if either the prediction or action model needs to be adapted for successful task completion.

## References

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344. https://doi.org/10.1038/nn.2921

Amiruddin, A., Fueggle, S. N., Nguyen, A. T., Gignac, G. E., Clunies-Ross, K. L., & Fox, A. M. (2017). Error monitoring and empathy: Explorations within a neurophysiological context. *Psychophysiology*, *54*(6), 864–873. https://doi.org/10.1111/psyp.12846

Amoruso, L., Sedeño, L., Huepe, D., Tomio, A., Kamienkowski, J., Hurtado, E., Cardona, J. F., Álvarez González, M. Á., Rieznik, A., Sigman, M., Manes, F., & Ibáñez, A. (2014). Time to Tango: Expertise and contextual anticipation during action observation. *NeuroImage*, *98*, 366–385. https://doi.org/10.1016/j.neuroimage.2014.05.005

Angelini, M., Fabbri-Destro, M., Lopomo, N. F., Gobbo, M., Rizzolatti, G., & Avanzini, P. (2018). Perspective-dependent reactivity of sensorimotor mu rhythm in alpha and beta ranges during action observation: an EEG study. *Scientific Reports*, *8*(1), 12429. https://doi.org/10.1038/s41598-018-30912-w

Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. *Cognition*, *106*(3), 1093–1108. https://doi.org/10.1016/j.cognition.2007.05.005

Barke, A., Bode, S., Dechent, P., Schmidt-Samoa, C., van Heer, C., & Stahl, J. (2017). To err is (perfectly) human: behavioural and neural correlates of error processing and perfectionism. *Social Cognitive and Affective Neuroscience*, *12*(10), 1647–1657. https://doi.org/10.1093/scan/nsx082

Barker, T. V., Troller-Renfree, S., Pine, D. S., & Fox, N. A. (2015). Individual differences in social anxiety affect the salience of errors in social contexts. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(4), 723–735. https://doi.org/10.3758/s13415-015-0360-9

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, *34*(2), 163-175. https://doi.org/10.1023/B:JADD.0000022607.19833.00

Bates, A. T., Kiehl, K. A., Laurens, K. R., & Liddle, P. F. (2002). Error-related negativity and correct response negativity in schizophrenia. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *113*(9), 1454–1463. https://doi.org/10.1016/S1388-2457(02)00154-2

Bates, A. T., Patel, T. P., & Liddle, P. F. (2005). External Behavior Monitoring Mirrors Internal Behavior Monitoring. *Journal of Psychophysiology*, *19*(4), 281–288. https://doi.org/10.1027/0269-8803.19.4.281

Becker, M. P. I., Nitsch, A. M., Miltner, W. H. R., & Straube, T. (2014). A Single-Trial Estimation of the Feedback-Related Negativity and Its Relation to BOLD Responses in a Time-Estimation Task. *The Journal of Neuroscience*, *34*(8), 3005. https://doi.org/10.1523/JNEUROSCI.3684-13.2014

Bellebaum, C., & Colosio, M. (2014). From Feedback- to Response-based Performance Monitoring in Active and Observational Learning. *Journal of Cognitive Neuroscience*, *26*(9), 2111–2127. https://doi.org/10.1162/jocn_a_00612

Bellebaum, C., Ghio, M., Wollmer, M., Weismüller, B., & Thoma, P. (2020). The role of trait empathy in the processing of observed actions in a false-belief task. *Social Cognitive and Affective Neuroscience*, *15.* https://doi.org/10.1093/scan/nsaa009

Bellebaum, C., Kobza, S., Thiele, S., & Daum, I. (2010). It Was Not MY Fault: Event-Related Brain Potentials in Active and Observational Learning from Feedback. *Cerebral Cortex (New York, N.Y.: 1991)*, *20*(12), 2874–2883. https://doi.org/10.1093/cercor/bhq038

Bench, C. J., Frith, C. D., Grasby, P. M., Friston, K. J., Paulesu, E., Frackowiak, R. S., & Dolan, R. J. (1993). Investigations of the functional anatomy of attention using the Stroop test. *Neuropsychologia*, *31*(9), 907–922. https://doi.org/10.1016/0028-3932(93)90147-r

Bernhardt, B. C., & Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience*, *35*, 1–23. https://doi.org/10.1146/annurev-neuro-062111-150536

Bernstein, P. S., Scheffers, M. K., & Coles, M. G. H. (1995). “Where did I go wrong?” A psychophysiological analysis of error detection. *Journal of Experimental Psychology. Human Perception and Performance*, *21*(6), 1312–1322. https://doi.org/10.1037//0096-1523.21.6.1312

Birch, S., & Bloom, P. (2007). The curse of knowledge in reasoning about false belief. *Psychological Science*, *18*, 382–386. https://doi.org/10.1111/j.1467-9280.2007.01909.x

Bonneville-Roussy, A., & Bouffard, T. (2014). When quantity is not enough: Disentangling the roles of practice time, self-regulation and deliberate practice in musical achievement. *Psychology of Music*, *43*(5), 686–704. https://doi.org/10.1177/0305735614534910

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (Eds.) (2001). *Conflict monitoring and cognitive control*.: Vol. *108*. American Psychological Association.

Brazil, I. A., Mars, R. B., Bulten, B. H., Buitelaar, J. K., Verkes, R. J., & de Bruijn, E. R. A. (2011). A neurophysiological dissociation between monitoring one’s own and others’ actions in psychopathy. *Biological Psychiatry*, *69*(7), 693–699. https://doi.org/10.1016/j.biopsych.2010.11.013

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Großhirnrinde*. Barth.

Brown, E. C., & Brüne, M. (2012). The role of prediction in social neuroscience. *Frontiers in Human Neuroscience*, *6*, 147. https://doi.org/10.3389/fnhum.2012.00147

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science (New York, N.Y.)*, *307*(5712), 1118–1121. https://doi.org/10.1126/science.1105783

Bufalari, I., Aprile, T., Avenanti, A., Di Russo, F., & Aglioti, S. M. (2007). Empathy for pain and touch in the human somatosensory cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(11), 2553–2561. https://doi.org/10.1093/cercor/bhl161

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *107*(32). https://doi.org/10.1073/pnas.1003111107

Burnside, R., Fischer, A. G., & Ullsperger, M. (2019). The feedback-related negativity indexes prediction error in active but not observational learning. *Psychophysiology*, *56*(9), e13389. https://doi.org/10.1111/psyp.13389

Buzzell, G. A., Beatty, P. J., Paquette, N. A., Roberts, D. M., & McDonald, C. G. (2017). Error-Induced Blindness: Error Detection Leads to Impaired Sensory Processing and Lower Accuracy at Short Response-Stimulus Intervals. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(11), 2895–2903. https://doi.org/10.1523/JNEUROSCI.1202-16.2017

Candidi, M., Maria Sacheli, L., Mega, I., & Aglioti, S. M. (2014). Somatotopic Mapping of Piano Fingering Errors in Sensorimotor Experts: TMS Studies in Pianists and Visually Trained Musically Naïves. *Cerebral Cortex (New York, N.Y.: 1991)*, *24*(2), 435–443. https://doi.org/10.1093/cercor/bhs325

Carp, J., Halenar, M. J., Quandt, L. C., Sklar, A., & Compton, R. J. (2009). Perceived similarity and neural mirroring: Evidence from vicarious error processing. *Social Neuroscience*, *4*(1), 85–96. https://doi.org/10.1080/17470910802083167

Carrasco, M., Harbin, S. M., Nienhuis, J. K., Fitzgerald, K. D., Gehring, W. J., & Hanna, G. L. (2013). Increased error-related brain activity in youth with obsessive-compulsive disorder and unaffected siblings. *Depression and Anxiety*, *30*(1), 39–46. https://doi.org/10.1002/da.22035

Carrasco, M., Hong, C., Nienhuis, J. K., Harbin, S. M., Fitzgerald, K. D., Gehring, W. J., & Hanna, G. L. (2013). Increased error-related brain activity in youth with obsessive-compulsive disorder and other anxiety disorders. *Neuroscience Letters*, *541*, 214–218. https://doi.org/10.1016/j.neulet.2013.02.017

Carter, C. S., Braver Todd S., Barch Deanna M., Botvinick Matthew M., Noll Douglas, & Cohen Jonathan D. (1998). Anterior Cingulate Cortex, Error Detection, and the Online Monitoring of Performance. *Science (New York, N.Y.)*, *280*(5364), 747–749. https://doi.org/10.1126/science.280.5364.747

Ceccarini, F., & Castiello, U. (2019). Your error in my hand: An investigation of observational posterror slowing. *Psychonomic Bulletin & Review*, *26*(1), 298–304. https://doi.org/10.3758/s13423-018-1495-9

Chang, A., Chen, C.-C., Li, H.-H., & Li, C.-S. R. (2014). Event-Related Potentials for Post-Error and Post-Conflict Slowing. *PloS One*, *9*(6), e99909. https://doi.org/10.1371/journal.pone.0099909

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience*, *23*(4), 936–946. https://doi.org/10.1162/jocn.2010.21456

Chiu, P. H., & Deldin, P. J. (2007). Neural Evidence for Enhanced Error Detection in Major Depressive Disorder. *American Journal of Psychiatry*, *164*(4), 608–616. https://doi.org/10.1176/ajp.2007.164.4.608

Clawson, A., Clayson, P. E., Worsham, W., Johnston, O., South, M., & Larson, M. J. (2014). How about watching others? Observation of error-related feedback by others in autism spectrum disorders. *International Journal of Psychophysiology*, *92*(1), 26-34. https://doi.org/10.1016/j.ijpsycho.2014.01.009

Cracco, E., Desmet, C., & Brass, M. (2016). When your error becomes my error: Anterior insula activation in response to observed errors is modulated by agency. *Social Cognitive and Affective Neuroscience*, *11*(3), 357–366. https://doi.org/10.1093/scan/nsv120

Crump, M. J. C., & Logan, G. D. (2013). Prevention and correction in post-error performance: An ounce of prevention, a pound of cure. *Journal of Experimental Psychology*, *142*(3), 692–709. https://doi.org/10.1037/a0030014

Damaso, K., Williams, P., & Heathcote, A. (2020). Evidence for different types of errors being associated with different types of post-error changes. *Psychonomic Bulletin & Review*, *27*(3), 435–440. https://doi.org/10.3758/s13423-019-01675-w

Danielmeier, C., Eichele, T., Forstmann, B. U., Tittgemeyer, M., & Ullsperger, M. (2011). Posterior Medial Frontal Cortex Activity Predicts Post-Error Adaptations in Task-Related Visual and Motor Areas. *Journal of Neuroscience*, *31*(5), 1780–1789. https://doi.org/10.1523/JNEUROSCI.4299-10.2011

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*(10), 85.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

de Bruijn, E. R. A., Lange, F. P. de, Cramon, D. Y. von, & Ullsperger, M. (2009). Where errors are rewarding. *The Journal of Neuroscience*, *29*(39), 12183–12186. https://doi.org/10.1523/JNEUROSCI.1751-09.2009

de Bruijn, E. R. A., & von Rhein, D. T. (2012). Is your error my concern? An event-related potential study on own and observed error detection in cooperation and competition. *Frontiers in Neuroscience*, *6*, 8. https://doi.org/10.3389/fnins.2012.00008

de Haen, J. (n.d.). *Deutsche Version der Cambridge Behavior Scale*. http://docs.autismresearchcentre.com/tests/EQ_Deutsch.pdf

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., Cramon, D. Y. von, & Engel, A. K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *25*(50), 11730–11737. https://doi.org/10.1523/JNEUROSCI.3286-05.2005

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a Neural System for Error Detection and Compensation. *Psychological Science*, *5*(5), 303–305. https://doi.org/10.1111/j.1467-9280.1994.tb00630.x

Desmet, C., Deschrijver, E., & Brass, M. (2014). How social is error observation? The neural mechanisms underlying the observation of human and machine errors. *Social Cognitive and Affective Neuroscience*, *9*(4), 427–435. https://doi.org/10.1093/scan/nst002

Devinsky, O., Morrell, M. J., & Vogt, B. A. (1995). Contributions of anterior cingulate cortex to behaviour. *Brain: A Journal of Neurology*, *118*(1), 279–306. https://doi.org/10.1093/brain/118.1.279

Di Gregorio, F., Maier, M. E., & Steinhauser, M. (2022). Early correlates of error-related
brain activity predict subjective timing of error awareness. *Psychophysiology*, *59*(7), e14020. https://doi.org/10.1111/psyp.14020

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, *91*(1), 176–180. https://doi.org/10.1007/BF00230027

Donnarumma, F., Costantini, M., Ambrosini, E., Friston, K., & Pezzulo, G. (2017). Action perception as hypothesis testing. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *89*, 45–60. https://doi.org/10.1016/j.cortex.2017.01.016

Drew, A. R., Quandt, L. C., & Marshall, P. J. (2015). Visual influences on sensorimotor EEG responses during observation of hand actions. *Brain Research*, *1597*, 119–128. https://doi.org/10.1016/j.brainres.2014.11.048

Endrass, T., Klawohn, J., Schuster, F., & Kathmann, N. (2008). Overactive performance monitoring in obsessive-compulsive disorder: ERP evidence from correct and erroneous reactions. *Neuropsychologia*, *46*(7), 1877–1887. https://doi.org/10.1016/j.neuropsychologia.2007.12.001

Endrass, T., Riesel, A., Kathmann, N., & Buhlmann, U. (2014). Performance monitoring in obsessive-compulsive disorder and social anxiety disorder. *Journal of Abnormal Psychology*, *123*(4), 705–714. https://doi.org/10.1037/abn0000012

Epley, N., Keysar, B., van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Eppinger, B., Kray, J., Mock, B., & Mecklinger, A. (2008). Better or worse than expected? Aging, learning, and the ERN. *Neuropsychologia*, *46*(2), 521–539. https://doi.org/10.1016/j.neuropsychologia.2007.09.001

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3). https://doi.org/10.1037/0033-295X.100.3.363

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences*, *15*(2), 85–93. https://doi.org/10.1016/j.tics.2010.11.004

Euser, A. S., Evans, B. E., Greaves-Lord, K., Huizink, A. C., & Franken, I. H. A. (2013). Diminished error-related brain activity as a promising endophenotype for substance-use disorders: evidence from high-risk offspring. *Addiction Biology*, *18*(6), 970–984. https://doi.org/10.1111/adb.12002

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. Ii. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*(6), 447–455. https://doi.org/10.1016/0013-4694(91)90062-9

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). Erp components on reaction errors and their functional significance: A tutorial. *Biological Psychology*, *51*(2-3), 87–107. https://doi.org/10.1016/s0301-0511(99)00031-9

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

Ferguson, H. J., Cane, J. E., Douchkov, M., & Wright, D. (2015). Empathy predicts false belief reasoning ability: Evidence from the N400. *Social Cognitive and Affective Neuroscience*, *10*(6), 848–855. https://doi.org/10.1093/scan/nsu131

Fiehler, K., Ullsperger, M., & von Cramon, D. Y. (2005). Electrophysiological correlates of error correction. *Psychophysiology*, *42*(1), 72–82. https://doi.org/10.1111/j.1469-8986.2005.00265.x

Fischer, A. G., & Ullsperger, M. (2013). Real and Fictive Outcomes Are Processed Differently but Converge on a Common Adaptive Mechanism. *Neuron*, *79*(6), 1243–1255. https://doi.org/10.1016/j.neuron.2013.07.006

Flanagan, J. R., & Johansson, R. S. (2003). Action plans used in action observation. *Nature*, *424*(6950), 769–771. https://doi.org/10.1038/nature01861

Forster, B., & Pavone, E. F. (2008). Electrophysiological correlates of crossmodal visual distractor congruency effects: Evidence for response conflict. *Cognitive, Affective & Behavioral Neuroscience*, *8*(1), 65–73. https://doi.org/10.3758/CABN.8.1.65

Franken, I. H. A., van Strien, J. W., Franzek, E. J., & van de Wetering, B. J. (2007). Error-processing deficits in patients with cocaine dependence. *Biological Psychology*, *75*(1), 45–51. https://doi.org/10.1016/j.biopsycho.2006.11.003

Frömer, R., Maier, M. E., & Rahman, R. A. (2018). Group-Level EEG-Processing Pipeline for Flexible Single Trial-Based Analyses Including Linear Mixed Models. *Frontiers in Neuroscience*, *12*, 48. https://doi.org/10.3389/fnins.2018.00048

Fu, Z., Wu, D.-A. J., Ross, I., Chung, J. M., Mamelak, A. N., Adolphs, R., & Rutishauser, U. (2019). Single-Neuron Correlates of Error Monitoring and

Post-Error Adjustments in Human Medial Frontal Cortex. *Neuron*, *101*(1), 165-177.e5. https://doi.org/10.1016/j.neuron.2018.11.016

Fukushima, H., & Hiraki, K. (2006). Perceiving an opponent's loss: Gender-related differences in the medial-frontal negativity. *Social Cognitive and Affective Neuroscience*, *1*(2), 149–157. https://doi.org/10.1093/scan/nsl020

Fukushima, H., & Hiraki, K. (2009). Whose loss is it? Human electrophysiological correlates of non-self reward processing. *Social Neuroscience*, *4*, 261–275. https://doi.org/10.1080/17470910802625009

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain: A Journal of Neurology*, *119 (Pt 2)*, 593–609. https://doi.org/10.1093/brain/119.2.593

Ganushchak, L. Y., & Schiller, N. O. (2008). Motivation and semantic context affect brain error-monitoring activity: An event-related brain potentials study. *NeuroImage*, *39*(1), 395–405. https://doi.org/10.1016/j.neuroimage.2007.09.001

Garavan, H., Ross, T. J., Murphy, K., Roche, R., & Stein, E. A. (2002). Dissociable Executive Functions in the Dynamic Control of Behavior: Inhibition, Error Detection, and Correction. *NeuroImage*, *17*(4), 1820–1829. https://doi.org/10.1006/nimg.2002.1326

Gawlowska, M., Domagalik, A., Beldzik, E., Marek, T., & Mojsa-Kaja, J. (2018). Dynamics of error-related activity in deterministic learning - an EEG and fMRI study. *Scientific Reports*, *8*(1), 14617. https://doi.org/10.1038/s41598-018-32995-x

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A Neural System for Error Detection and Compensation. *Psychological Science*, *4*(6), 385–390.

Gehring, W. J., Himle, J., & Nisenson, L. G. (2000). Action-monitoring dysfunction in obsessive-compulsive disorder. *Psychological Science*, *11*(1), 1–6. https://doi.org/10.1111/1467-9280.00206

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (Eds.). (2012). *The Oxford Handbook of Event-Related Potential*. *The Error-Related Negativity (ERN/Ne)*. Oxford University Press.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science (New York, N.Y.)*, *295*(5563), 2279–2282. https://doi.org/10.1126/science.1066893

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

Gujing, L., Hui, H., Xin, L., Lirong, Z., Yutong, Y., Guofeng, Y., Jing, L., Shulin, Z., Lei, Y., Cheng, L., & Dezhong, Y. (2019). Increased Insular Connectivity and Enhanced Empathic Ability Associated with Dance/Music Training. *Neural Plasticity*, *2019*, 9693109. https://doi.org/10.1155/2019/9693109

Hajcak, G., McDonald, N., & Simons, R. F. (2003). To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behavior. *Psychophysiology*, *40*(6), 895–903. https://doi.org/10.1111/1469-8986.00107

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., Yeung, N., & Simons, R. F. (2005). On the ERN and the significance of errors. *Psychophysiology*, *42*(2), 151–160. https://doi.org/10.1111/j.1469-8986.2005.00270.x

Hajcak, G., & Simons, R. F. (2008). Oops! I did it again: An ERP and behavioral study of double-errors. *Brain and Cognition*, *68*(1), 15–21. https://doi.org/10.1016/j.bandc.2008.02.118

Hall, J. A., & Schwartz, R. (2019). Empathy present and future. *The Journal of Social Psychology*, *159*(3), 225–243. https://doi.org/10.1080/00224545.2018.1477442

Harris, L. N., Perfetti, C. A., & Rickles, B. (2014). Error-related negativities during spelling judgments expose orthographic knowledge. *Neuropsychologia*, *54*, 112–128. https://doi.org/10.1016/j.neuropsychologia.2013.12.007

Herrojo Ruiz, M., Jabusch, H.-C., & Altenmüller, E. (2009). Detecting wrong notes in advance: Neuronal correlates of error monitoring in pianists. *Cerebral Cortex (New York, N.Y.: 1991)*, *19*(11), 2625–2639. https://doi.org/10.1093/cercor/bhp021

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*(4), 304–309. https://doi.org/10.1038/1124

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., Dien, J., & Coles, M. G. (1998). Error-related scalp potentials elicited by hand and foot movements: Evidence for an output-independent error-processing system in humans. *Neuroscience Letters*, *242*(2), 65–68. https://doi.org/10.1016/s0304-3940(98)00035-4

Holroyd, C. B., Hajcak, G., & Larsen, J. T. (2006). The good, the bad and the neutral: Electrophysiological responses to feedback stimuli. *Brain Research*, *1105*(1), 93–101. https://doi.org/10.1016/j.brainres.2005.12.015

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G. H., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*(5), 497–498. https://doi.org/10.1038/nn1238

Houtman, F., Castellar, E. N., & Notebaert, W. (2012). Orienting to errors with and without immediate feedback. *Journal of Cognitive Psychology*, *24*(3), 278–285. https://doi.org/10.1080/20445911.2011.617301

Itagaki, S., & Katayama, J. (2008). Self-relevant criteria determine the evaluation of outcomes induced by others. *Neuroreport*, *19*(3). https://doi.org/10.1097/wnr.0b013e3282f556e8

Jentzsch, I., & Leuthold, H. (2006). Short article: Control over speeded actions: A common processing locus for micro- and macro-trade-offs? *Quarterly Journal of Experimental Psychology*, *59*(8), 1329–1337. https://doi.org/10.1080/17470210600674394

Jentzsch, I., Mkrtchian, A., & Kansal, N. (2014). Improved effectiveness of performance monitoring in amateur instrumental musicians. *Neuropsychologia*, *52*, 117–124. https://doi.org/10.1016/j.neuropsychologia.2013.09.025

Jessup, R. K., Busemeyer, J. R., & Brown, J. W. (2010). Error effects in anterior cingulate cortex reverse when error likelihood is high. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(9), 3467–3472. https://doi.org/10.1523/JNEUROSCI.4130-09.2010

Jones, A. P., Happé, F. G., Gilbert, F., Burnett, S., & Viding, E. (2010). Feeling, caring, knowing: different types of empathy deficit in boys with psychopathic tendencies and autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, *51*(11), 1188–1197. https://doi.org/10.1111/j.1469-7610.2010.02280.x

Kalfaoğlu, Ç., & Stafford, T. (2014). Performance Breakdown Effects Dissociate from Error Detection Effects in Typing. *Quarterly Journal of Experimental Psychology*, *67*(3), 508-524. https://doi.org/10.1080/17470218.2013.820762

Kalfaoğlu, Ç., Stafford, T., & Milne, E. (2018). Frontal theta band oscillations predict error correction and posterror slowing in typing. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(1), 69. https://doi.org/10.1037/xhp0000417

Kang, S. K., Hirsh, J. B., & Chasteen, A. L. (2010). Your mistakes are mine: Self-other overlap predicts neural response to observed errors. *Journal of*

*Experimental Social Psychology*, *46*(1), 229–232. https://doi.org/10.1016/j.jesp.2009.09.012

Kiehl, K. A., Liddle, P. F., & Hopfinger, J. B. (2000). Error processing and the rostral anterior cingulate: An event-related fMRI study. *Psychophysiology*, *37*(2), 216–223. https://doi.org/10.1111/1469-8986.3720216

Kim, E. Y., Iwaki, N., Uno, H., & Fujita, T. (2005). Error-Related Negativity in Children: Effect of an Observer. *Developmental Neuropsychology*, *28*(3), 871–883. https://doi.org/10.1207/s15326942dn2803_7

Koban, L., & Pourtois, G. (2014). Brain systems underlying the affective and social monitoring of actions: An integrative review. *Neuroscience and Biobehavioral Reviews*, *46 Pt 1*, 71–84. https://doi.org/10.1016/j.neubiorev.2014.02.014

Koban, L., Pourtois, G., Bediou, B., & Vuilleumier, P. (2012). Effects of social context and predictive relevance on action outcome monitoring. *Cognitive, Affective & Behavioral Neuroscience*, *12*(3), 460–478. https://doi.org/10.3758/s13415-012-0091-0

Koban, L., Pourtois, G., Vocat, R., & Vuilleumier, P. (2010). When your errors make me lose or win: Event-related potentials to observed errors of cooperators and competitors. *Social Neuroscience*, *5*, 360–374. https://doi.org/10.1080/17470911003651547

Kobza, S., & Bellebaum, C. (2013). Mediofrontal event-related potentials following observed actions reflect an action prediction error. *The European Journal of Neuroscience*, *37*(9), 1435–1440. https://doi.org/10.1111/ejn.12138

Kobza, S., Thoma, P., Daum, I., & Bellebaum, C. (2011). The feedback-related negativity is modulated by feedback probability in observational learning. *Behavioural Brain Research*, *225*(2), 396–404. https://doi.org/10.1016/j.bbr.2011.07.059

Krigolson, O. E., Hassall, C. D., & Handy, T. C. (2014). How We Learn to Make Decisions: Rapid Propagation of Reinforcement Learning Prediction Errors in Humans. *Journal of Cognitive Neuroscience*, *26*(3), 635–644. https://doi.org/10.1162/jocn_a_00509

Kruse-Weber, S., & Parncutt, R. (2014). Error management for musicians: An interdisciplinary conceptual framework. *Frontiers in Psychology*, *5*, 777. https://doi.org/10.3389/fpsyg.2014.00777

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

Ladouceur, C. D., Slifka, J. S., Dahl, R. E., Birmaher, B., Axelson, D. A., & Ryan, N. D. (2012). Altered error-related brain activity in youth with major depression. *Developmental Cognitive Neuroscience*, *2*(3), 351–362. https://doi.org/10.1016/j.dcn.2012.01.005

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

Lamm, C., Meltzoff, A. N., & Decety, J. (2010). How do we empathize with someone who is not like us? A functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, *22*(2), 362–376. https://doi.org/10.1162/jocn.2009.21186

Large, E. W. (1993). Dynamic programming for the analysis of serial behaviors. *Behavior Research Methods, Instruments & Computers*, *25*(2). https://doi.org/10.3758/BF03204504

Larson, M. J., Fair, J. E., Good, D. A., & Baldwin, S. A. (2010). Empathy and error processing. *Psychophysiology*, *47*(3), 415–424. https://doi.org/10.1111/j.1469-8986.2009.00949.x

Li, Y., & Feng, T. (2020). The effects of sport expertise and shot results on basketball players' action anticipation. *PloS One*, *15*(1), e0227521. https://doi.org/10.1371/journal.pone.0227521

Littel, M., van den Berg, I., Luijten, M., van Rooij, A. J., Keemink, L., & Franken, I. H. A. (2012). Error processing and response inhibition in excessive computer game players: an event-related potential study. *Addiction Biology*, *17*(5), 934–947. https://doi.org/10.1111/j.1369-1600.2012.00467.x

Lockwood, P. L., Apps, M. A. J., Roiser, J. P., & Viding, E. (2015). Encoding of Vicarious Reward Prediction in Anterior Cingulate Cortex and Relationship with Trait Empathy. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(40), 13720–13727. https://doi.org/10.1523/JNEUROSCI.1703-15.2015

Loehr, J. D., Kourtis, D., Vesper, C., Sebanz, N., & Knoblich, G. (2013). Monitoring Individual and Joint Action Outcomes in Duet Music Performance. *Journal of Cognitive Neuroscience*, *25*(7), 1049–1061. https://doi.org/10.1162/jocn_a_00388

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

Luijten, M., van Meel, C. S., & Franken, I. H. (2011). Diminished error processing in smokers during smoking cue exposure. *Pharmacology Biochemistry and Behavior*, *97*(3), 514–520. https://doi.org/10.1016/j.pbb.2010.10.012

Maidhof, C., Pitkäniemi, A., & Tervaniemi, M. (2013). Predictive error detection in pianists: A combined ERP and motion capture study. *Frontiers in Human Neuroscience*, *7*, 587. https://doi.org/10.3389/fnhum.2013.00587

Maidhof, C., Rieger, M., Prinz, W., & Koelsch, S. (2009). Nobody is perfect: Erp effects prior to performance errors in musicians indicate fast monitoring processes. *PloS One*, *4*(4), e5032. https://doi.org/10.1371/journal.pone.0005032

Maier, M. E., Di Pellegrino, G., & Steinhauser, M. (2012). Enhanced error-related negativity on flanker errors: Error expectancy or error significance? *Psychophysiology*, *49*(7), 899–908. https://doi.org/10.1111/j.1469-8986.2012.01373.x

Maier, M. E., & Steinhauser, M. (2016). Error significance but not error expectancy predicts error-related negativities for different error types. *Behavioural Brain Research*, *297*, 259–267. https://doi.org/10.1016/j.bbr.2015.10.031

Maier, M. E., Steinhauser, M., & Hübner, R. (2008). Is the error-related negativity amplitude related to error detectability? Evidence from effects of different error types. *Journal of Cognitive Neuroscience*, *20*(12), 2263–2273. https://doi.org/10.1162/jocn.2008.20159

Marco-Pallarés, J., Camara, E., Münte, T. F., & Rodríguez-Fornells, A. (2008). Neural Mechanisms Underlying Adaptive Actions after Slips. *Journal of Cognitive Neuroscience*, *20*(9), 1595–1610. https://doi.org/10.1162/jocn.2008.20117

Marco-Pallarés, J., Krämer, U. M., Strehl, S., Schröder, A., & Münte, T. F. (2010). When decisions of others matter to me: An electrophysiological analysis. *BMC Neuroscience*, *11*, 86. https://doi.org/10.1186/1471-2202-11-86

Marhe, R., van de Wetering, B. J., & Franken, I. H. (2013). Error-Related Brain Activity Predicts Cocaine Use After Treatment at 3-Month Follow-up. *Biological Psychiatry*, *73*(8), 782–788. https://doi.org/10.1016/j.biopsych.2012.12.016

Marshall, J. D., Li, T., Wu, J. H., & Dunn, T. W. (2022). Leaving flatland: Advances in 3D behavioral measurement. *Current Opinion in Neurobiology*, *73*, 102522. https://doi.org/10.1016/j.conb.2022.02.002

Mathalon, D. H., Whitfield, S. L., & Ford, J. M. (2003). Anatomy of an error: ERP and fMRI. *Biological Psychology*, *64*(1), 119–141. https://doi.org/10.1016/S0301-0511(03)00105-4

Meinz, E. J., & Hambrick, D. Z. (2010). Deliberate Practice Is Necessary but Not Sufficient to Explain Individual Differences in Piano Sight-Reading Skill: The Role of Working Memory Capacity. *Psychological Science*, *21*(7), 914–919. https://doi.org/10.1177/0956797610373933

Meyer, M. L., & Lieberman, M. (2012). Social Working Memory: Neurocognitive Networks and Directions for Future Research. *Frontiers in Psychology*, *3*, 571. https://doi.org/10.3389/fpsyg.2012.00571

Meyer, M. L., Spunt, R. P., Berkman, E. T., Taylor, S. E., & Lieberman, M. D. (2012). Evidence for social working memory from a parametric functional MRI study. *Proceedings of the National Academy of Sciences*, *109*(6), 1883. https://doi.org/10.1073/pnas.1121077109

Miltner, W. H. R., Brauer, J., Hecht, H., Trippe, R., & Coles, M. (2004). Parallel brain activity for self-generated and observed errors. In *Errors, Conflicts, and the Brain. Current Opinions on Performance Monitoring* (pp. 124-129). Max Planck Institute for Human Cognitive and Brain Sciences.

Miltner, W. H., Braun, C. H., & Coles, M. G. (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

Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A. J., Schweizer, S., Frith, C. D., & Dalgleish, T. (2009). A key role for similarity in vicarious reward. *Science (New York, N.Y.)*, *324*(5929), 900. https://doi.org/10.1126/science.1170539

Morris, S. E., Yee, C. M., & Nuechterlein, K. H. (2006). Electrophysiological analysis of error monitoring in schizophrenia. *Journal of Abnormal Psychology*, *115*(2), 239. https://doi.org/10.1037/0021-843X.115.2.239

Morsel, A. M., Morrens, M., Temmerman, A., Sabbe, B., & de Bruijn, E. R. A. (2014). Electrophysiological (EEG) evidence for reduced performance monitoring in euthymic bipolar disorder. *Bipolar Disorders*, *16*(8), 820–829. https://doi.org/10.1111/bdi.12256

Murata, A., & Katayama, J. (2005). An unnecessary response is detected faster than an insufficient response. *Neuroreport*, *16*(14), 1595–1598. https://doi.org/10.1097/01.wnr.0000179080.62529.0b

Neta, M., Miezin, F. M., Nelson, S. M., Dubis, J. W., Dosenbach, N. U., Schlaggar, B. L., & Petersen, S. E. (2015). Spatial and Temporal Characteristics of Error-Related Activity in the Human Brain. *The Journal of Neuroscience*, *35*(1), 253. https://doi.org/10.1523/JNEUROSCI.1313-14.2015

Newman-Norlund, R. D., Ganesh, S., van Schie, H. T., Bruijn, E. R. A. de, & Bekkering, H. (2009). Self-identification and empathy modulate error-related brain activity during the observation of penalty shots between friend and foe. *Social Cognitive and Affective Neuroscience*, *4*(1), 10–22. https://doi.org/10.1093/scan/nsn028

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. H. (2004). Reinforcement-related brain potentials from medial frontal cortex: Origins and functional significance. *Neuroscience and Biobehavioral Reviews*, *28*(4), 441–448. https://doi.org/10.1016/j.neubiorev.2004.05.003

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, *38*(5), 752–760. https://doi.org/10.1111/1469-8986.3850752

Nieuwenhuis, S., Yeung, N., van den Wildenberg, W., & Ridderinkhof, K. R. (2003). Electrophysiological correlates of anterior cingulate function in a go/no-go task: Effects of response conflict and trial type frequency. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(1), 17–26. https://doi.org/10.3758/CABN.3.1.17

Ninomiya, T., Noritake, A., Ullsperger, M., & Isoda, M. (2018). Performance monitoring in the medial frontal cortex and related neural networks: From monitoring self actions to understanding others' actions. *Neuroscience Research*, *137*, 1–10. https://doi.org/10.1016/j.neures.2018.04.004

Notebaert, W., Houtman, F., van Opstal, F., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: An orienting account. *Cognition*, *111*(2), 275–279. https://doi.org/10.1016/j.cognition.2009.02.002

Núñez Castellar, E., Kühn, S., Fias, W., & Notebaert, W. (2010). Outcome expectancy and not accuracy determines posterror slowing: Erp support. *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 270–278. https://doi.org/10.3758/CABN.10.2.270

Núñez Castellar, E., Notebaert, W., van den Bossche, L., & Fias, W. (2011). How Monitoring Other's Actions Influences One's Own Performance. *Experimental Psychology*, *58*(6), 499–508. https://doi.org/10.1027/1618-3169/a000118

O'Doherty, J., Critchley, H., Deichmann, R., & Dolan, R. J. (2003). Dissociating Valence of Outcome from Behavioral Control in Human Orbital and Ventral

Prefrontal Cortices. *The Journal of Neuroscience*, *23*(21), 7931. https://doi.org/10.1523/JNEUROSCI.23-21-07931.2003

Oliveira, F. T. P., McDonald, J. J., & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error related: Expectancy deviation and the representation of action-outcome associations. *Journal of Cognitive Neuroscience*, *19*(12), 1994–2004. https://doi.org/10.1162/jocn.2007.19.12.1994

Özkan, D. G., Pezzetta, R., Moreau, Q., Abreu, A. M., & Aglioti, S. M. (2019). Predicting the fate of basketball throws: an EEG study on expert action prediction in wheelchair basketball players. *Experimental Brain Research*, *237*(12), 3363–3373. https://doi.org/10.1007/s00221-019-05677-x

Paas, A., Novembre, G., Lappe, C., & Keller, P. E. (2021). Not all errors are alike: Modulation of error-related neural responses in musical joint action. *Social Cognitive and Affective Neuroscience*, *16*(5), 512–524. https://doi.org/10.1093/scan/nsab019

Palmer, C., & Drake, C. (1997). Monitoring and planning capacities in the acquisition of music performance skills. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, *51*(4). https://doi.org/10.1037/1196-1961.51.4.369

Palmer, C., & van de Sande, C. (1993). Units of knowledge in music performance. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *19*(2), 457–470. https://doi.org/10.1037//0278-7393.19.2.457

Panasiti, M. S., Pavone, E. F., & Aglioti, S. M. (2016). Electrocortical signatures of detecting errors in the actions of others: An EEG study in pianists, non-pianist musicians and musically naïve people. *Neuroscience*, *318*, 104–113. https://doi.org/10.1016/j.neuroscience.2016.01.023

Perrone-McGovern, K., Simon-Dack, S., Esche, A., Thomas, C., Beduna, K., Rider, K., Spurling, A., & Matsen, J. (2017). The influence of emotional intelligence and perfectionism on Error-Related Negativity: An event related potential study. *Personality and Individual Differences*, *111*, 65–70. https://doi.org/10.1016/j.paid.2017.02.009

Pezzetta, R., Nicolardi, V., Tidoni, E., & Aglioti, S. M. (2018). Error, rather than its probability, elicits specific electrocortical signatures: A combined EEG-immersive virtual reality study of action observation. *Journal of Neurophysiology*, *120*(3), 1107–1118. https://doi.org/10.1152/jn.00130.2018

Pietschmann, M., Simon, K., Endrass, T., & Kathmann, N. (2008). Changes of performance monitoring with learning in older and younger adults. *Psychophysiology*, *45*(4), 559–568. https://doi.org/10.1111/j.1469-8986.2008.00651.x

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

Rabbitt, P. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, *71*(2), 264–272. https://doi.org/10.1037/h0022853

Rabbitt, P. (1969). Psychological refractory delay and response-stimulus interval duration in serial, choice-response tasks. *Acta Psychologica*, *30*, 195–219. https://doi.org/10.1016/0001-6918(69)90051-1

Rabinowitch, T.-C., Cross, I., & Burnard, P. (2012). Long-term musical group interaction has a positive influence on empathy in children. *Psychology of Music*, *41*, 484–498. https://doi.org/10.1177/0305735612440609

Rachaveti, D., Ranganathan, R., & SKM, V. (2020). Practice modifies the response to errors during a novel motor sequence learning task. *BioRxiv*, 2020.10.09.334169. https://doi.org/10.1101/2020.10.09.334169

Rankin, S. K., Large, E. W., & Fink, P. W. (2009). Fractal Tempo Fluctuation and Pulse Prediction. *Music Perception*, *26*(5), 401–413. https://doi.org/10.1525/mp.2009.26.5.401

Reinhart, R. M. G., Carlisle, N. B., Kang, M.-S., & Woodman, G. F. (2012). Event-related potentials elicited by errors during the stop-signal task. Ii: Human effector-specific error responses. *Journal of Neurophysiology*, *107*(10), 2794–2807. https://doi.org/10.1152/jn.00803.2011

Reniers, R., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A Questionnaire of Cognitive and Affective Empathy. *Journal of Personality Assessment*, *93*(1), 84–95. https://doi.org/10.1080/00223891.2010.528484

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science (New York, N.Y.)*, *306*(5695), 443–447. https://doi.org/10.1126/science.1100301

Ridderinkhof, K. R., van den Wildenberg, W. P., Segalowitz, S. J., & Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: The role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain and Cognition*, *56*(2), 129–140. https://doi.org/10.1016/j.bandc.2004.09.016

Ruchsow, M., Herrnberger, B., Wiesend, C., Grön, G., Spitzer, M., & Kiefer, M. (2004). The effect of erroneous responses on response monitoring in patients with major depressive disorder: A study with event-related potentials. *Psychophysiology*, *41*(6), 833–840. https://doi.org/10.1111/j.1469-8986.2004.00237.x

Santesso, D. L., & Segalowitz, S. J. (2009). The error-related negativity is related to risk taking and empathy in young men. *Psychophysiology*, *46*(1), 143–152. https://doi.org/10.1111/j.1469-8986.2008.00714.x

Scangos, K. W., Aronberg, R., & Stuphorn, V. (2013). Performance monitoring by presupplementary and supplementary motor area during an arm movement countermanding task. *Journal of Neurophysiology*, *109*(7), 1928–1939. https://doi.org/10.1152/jn.00688.2012

Scheffers, M. K., & Coles, M. G. H. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology. Human Perception and Performance*, *26*(1), 141–151. https://doi.org/10.1037//0096-1523.26.1.141

Schiffer, A.-M., Krause, K. H., & Schubotz, R. I. (2014). Surprisingly correct: Unexpectedness of observed actions activates the medial prefrontal cortex. *Human Brain Mapping*, *35*(4), 1615–1629. https://doi.org/10.1002/hbm.22277

Schuch, S., & Tipper, S. P. (2007). On observing another person's actions: Influences of observed inhibition and errors. *Perception & Psychophysics*, *69*(5), 828–837. https://doi.org/10.3758/BF03193782

Schultz, W., Dayan, P, & Montague P. R. (1997). A Neural Substrate of Prediction and Reward. *Science (New York, N.Y.)*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Shane, M. S., Stevens, M., Harenski, C. L., & Kiehl, K. A. (2008). Neural correlates of the processing of another's mistakes: A possible underpinning for social and observational learning. *NeuroImage*, *42*(1), 450–459. https://doi.org/10.1016/j.neuroimage.2007.12.067

Shane, M. S., Stevens, M. C., Harenski, C. L., & Kiehl, K. A. (2009). Double dissociation between perspective-taking and empathic-concern as predictors of hemodynamic response to another's mistakes. *Social Cognitive and Affective Neuroscience*, *4*(2), 111–118. https://doi.org/10.1093/scan/nsn043

Shohamy, D., Myers, C. E., Kalanithi, J., & Gluck, M. A. (2008). Basal ganglia and dopamine contributions to probabilistic category learning. *Neuroscience & Biobehavioral Reviews*, *32*(2), 219–236. https://doi.org/10.1016/j.neubiorev.2007.07.008

Simmonite, M., Bates, A. T., Groom, M. J., Jackson, G. M., Hollis, C., & Liddle, P. F. (2012). Error processing-associated event-related potentials in schizophrenia and unaffected siblings. *International Journal of Psychophysiology: Official Journal of the International Organization of*

*Psychophysiology*, *84*(1), 74–79. https://doi.org/10.1016/j.ijpsycho.2012.01.012

Singer, T., Seymour, B., O’Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science (New York, N.Y.)*, *303*(5661), 1157–1162. https://doi.org/10.1126/science.1093535

Singer, T., Seymour, B., O’Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*(7075), 466–469. https://doi.org/10.1038/nature04271

Spinnato, J., Roubaud, M.-C., Burle, B., & Torrésani, B. (2015). Detecting single-trial EEG evoked potential using a wavelet domain linear mixed model: Application to error potentials classification. *Journal of Neural Engineering*, *12*(3), 36013. https://doi.org/10.1088/1741-2560/12/3/036013

Stahl, J., Acharki, M., Kresimon, M., Völler, F., & Gibbons, H. (2015). Perfect error processing: Perfectionism-related variations in action monitoring and error processing mechanisms. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *97*(2), 153–162. https://doi.org/10.1016/j.ijpsycho.2015.06.002

Steinhauser, M., & Andersen, S. K. (2019). Rapid adaptive adjustments of selective attention following errors revealed by the time course of steady-state visual evoked potentials. *NeuroImage*, *186*, 83–92. https://doi.org/10.1016/j.neuroimage.2018.10.059

Stevens, F. L., Hurley, R. A., Taber, K. H., & Hayman, L. A. (2011). Anterior Cingulate Cortex: Unique Role in Cognition and Emotion. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *23*(2), 121–125. https://doi.org/10.1176/jnp.23.2.jnp121

Stewart, A. X., Nuthmann, A., & Sanguinetti, G. (2014). Single-trial classification of EEG in a visual object task using ICA and machine learning. *Journal of Neuroscience Methods*, *228*, 1–14. https://doi.org/10.1016/j.jneumeth.2014.02.014

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Current Biology: CB*, *28*(19), 3128-3135.e8. https://doi.org/10.1016/j.cub.2018.07.052

Taylor, S. F., Stern, E. R., & Gehring, W. J. (2007). Neural systems for error monitoring: Recent findings and theoretical perspectives. *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *13*(2), 160–172. https://doi.org/10.1177/1073858406298184

Thompson-Schill, S. L., D’Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(26), 14792–14797. https://doi.org/10.1073/pnas.94.26.14792

Tucker, D. M., Liotti, M., Potts, G. F., Russell, G. S., & Posner, M. I. (1993). Spatiotemporal analysis of brain electrical fields. *Human Brain Mapping*, *1*(2), 134–152. https://doi.org/10.1002/hbm.460010206

Ullsperger, M., & Cramon, D. (2004). Neuroimaging of Performance Monitoring: Error Detection and Beyond. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *40*, 593–604. https://doi.org/10.1016/S0010-9452(08)70155-2

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

Urai, A. E., de Gee, J. W., Tsetsos, K., & Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. *ELife*, *8.* https://doi.org/10.7554/eLife.46331

van ’t Ent, D., & Apkarian, P. (1999). Motoric response inhibition in finger movement and saccadic eye movement: a comparative study. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *110*(6), 1058–1072. https://doi.org/10.1016/S1388-2457(98)00036-4

van Meel, C. S., & van Heijningen, C. A. A. (2010). The effect of interpersonal competition on monitoring internal and external error feedback. *Psychophysiology*, *47*(2), 213–222. https://doi.org/10.1111/j.1469-8986.2009.00944.x

van Schie, H. T., Mars, R. B., Coles, M. G. H., & Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, *7*(5), 549–554. https://doi.org/10.1038/nn1239

van Veen, V., & Carter, C. S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience*, *14*(4), 593–602. https://doi.org/10.1162/08989290260045837

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The Musical Ear Test, a new reliable test for measuring musical competence. *Learning and Individual Differences*, *20*(3), 188–196. https://doi.org/10.1016/j.lindif.2010.02.004

Wang, L., Tang, D., Zhao, Y., Hitchman, G., Wu, S., Tan, J., & Chen, A. (2015). Disentangling the impacts of outcome valence and outcome frequency on the post-error slowing. *Scientific Reports*, *5*, 8708. https://doi.org/10.1038/srep08708

Weinberg, A., Olvet, D. M., & Hajcak, G. (2010). Increased error-related brain activity in generalized anxiety disorder. *Biological Psychology*, *85*(3), 472–480. https://doi.org/10.1016/j.biopsycho.2010.09.011

Wellman, H., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development*, *72*, 655–684. https://doi.org/10.1111/1467-8624.00304

Wessel, J. R. (2018). An adaptive orienting theory of error processing. *Psychophysiology*, *55*(3). https://doi.org/10.1111/psyp.13041

Wessel, J. R., Danielmeier, C., Morton, J. B., & Ullsperger, M. (2012). Surprise and error: Common neuronal architecture for the processing of errors and novelty. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(22), 7528–7537. https://doi.org/10.1523/JNEUROSCI.6352-11.2012

Wessel, J. R., Klein, T. A., Ott, D. V., & Ullsperger, M. (2014). Lesions to the prefrontal performance-monitoring network disrupt neural processing and adaptive behaviors after both errors and novelty. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *50*, 45–54. https://doi.org/10.1016/j.cortex.2013.09.002

Williamon, A., & Valentine, E. (2000). Quantity and quality of musical practice as predictors of performance quality. *British Journal of Psychology*, *91*(3), 353–376. https://doi.org/10.1348/000712600161871

Williams, P., Heathcote, A., Nesbitt, K., & Eidels, A. (2016). Post-error recklessness and the hot hand. *Judgment and Decision Making*, *11*(2), 174-184.

Wirth, C., Lacey, E., Dockree, P., & Arvaneh, M. (2018). Single-Trial EEG Classification of Similar Errors. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2018*, 1919–1922. https://doi.org/10.1109/EMBC.2018.8512700

Wittfoth, M., Küstermann, E., Fahle, M., & Herrmann, M. (2008). The influence of response conflict on error processing: Evidence from event-related fMRI. *Brain Research*, *1194*, 118–129. https://doi.org/10.1016/j.brainres.2007.11.067

Wright, J. K., Grainger, S. A., Coundouris, S. P., & Henry, J. D. (2021). Affective empathy in neurodegenerative disorders: The importance of measurement type. *Neuroscience & Biobehavioral Reviews*, *127*, 808–819. https://doi.org/10.1016/j.neubiorev.2021.05.020

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931–959. https://doi.org/10.1037/0033-295x.111.4.939

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP Correlates of Feedback and Reward Processing in the Presence and Absence of Response Choice. *Cerebral Cortex (New York, N.Y.: 1991)*, *15*(5), 535–544. https://doi.org/10.1093/cercor/bhh153

Yoshida, K., Saito, N., Iriki, A., & Isoda, M. (2012). Social error monitoring in macaque frontal cortex. *Nature Neuroscience*, *15*(9), 1307–1312. https://doi.org/10.1038/nn.3180

Yu, R., & Zhou, X. (2006). Brain responses to outcomes of one's own and other's performance in a gambling task. *Neuroreport*, *17*(16). https://doi.org/10.1097/01.wnr.0000239960.98813.50

Zhao, Q., Wang, Y., Chen, Y., Wang, Y., Zhou, C., & Lu, Y. (2021). Expertise influences congruency monitoring during action observation at the motor level. *Social Cognitive and Affective Neuroscience*, *16*(12), 1288–1298. https://doi.org/10.1093/scan/nsab078

Zubarev, I., & Parkkonen, L. (2018). Evidence for a general performance-monitoring system in the human brain. *Human Brain Mapping*, *39*(11), 4322–4333. https://doi.org/10.1002/hbm.24273

# Affidavit

Eidesstattliche Erklärung gemäß § 5 der Promotionsordnung vom 15.06.2018 der Mathematisch-Naturwissenschaftlichen Fakultät der Heinrich-Heine-Universität Düsseldorf:

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist. Die Dissertation wurde in der vorliegenden oder ähnlichen Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den ______________ ________________

Datum Christine Albrecht

## Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Dr. Christian Bellebaum for teaching, guiding and supporting me throughout the last four years. Special thanks for always keeping your door open for any questions or problems I had and answering them with so much patience.

I am very grateful that I could share my PhD journey, my office, all the ups and downs and every random thought with Alexander Seidel. Thank you for bearing with me, keeping me up to date with nerd-culture and for taking off those noise-cancelling headphones every time I threw my rubber at you. Thank you, Laura Bechtold, for lunch hours and dog walks filled with work-related and -unrelated conversations, for mental health checks and for the best time-management tips. Thank you, Marta Ghio, for believing in my scientific abilities all these years ago and for always offering a kind word or advice when it was needed. Thank you to all my other friends and colleagues from the department of Biological Psychology for making the last years an exciting and amazing experience.

Thank you to Karolina Kaczmarczyk and Marco Jahn, my home office coworkers, for helping me to keep sane and for somehow managing to create new, wonderful memories even during a pandemic.

Thank you to my family, my friends and my partner for always having my back and reminding me about all the great people and activities that can be found outside the office.

Lastly, thank you to my students for offering me new perspectives on my research, and to everyone who participated in, advertised or conducted the experiments.

# Appendix

### Original article of Study 1

Albrecht, C., & Bellebaum, C. (2021). Effects of trait empathy and expectation on the processing of observed actions. Cognitive, affective & behavioral neuroscience, 21(1), 156–171. https://doi.org/10.3758/s13415-020-00857-7

I was the main author of this article. I designed the conceptualization and methodology, and developed the extension of the original paradigm by Kobza & Bellebaum (2020) together with Christian Bellebaum. I programmed the extension of the original paradigm by Kobza & Bellebaum (2013). I planned and supervised the data acquisition, analyzed, and interpreted the data.

### Original article of Study 2

Albrecht, C., & Bellebaum, C. (2021). Disentangling effects of expectancy, accuracy, and empathy on the processing of observed actions. *Psychophysiology*, *58*(9), e13883. https://doi.org/10.1111/psyp.13883

I was the main author of this article. A paradigm developed by Kobza & Bellebaum (2013) and extended by Bellebaum et al. (2020) was used for this study. Half of the sample was acquired for the Bellebaum et al. (2020) study by Christian Bellebaum and colleagues. The data acquisition of the other half of the sample was planned and supervised by Christian Bellebaum. I developed the research question and subsequent analysis method used in the study. I further analyzed and interpreted the data.

### Original manuscript of Study 3

Albrecht, C., & Bellebaum, C. (2022). Slip or fallacy? Effects of error type on own and observed pitch error processing in pianists. *Manuscript submitted for publication at Cognitive, affective & behavioral neuroscience.*

I was the main author of this manuscript. I designed the conceptualization and methodology together with Christian Bellebaum. I further developed and programmed the paradigm for both experiments. I planned, supervised and partly conducted the data acquisition. I analyzed and interpreted the data.

Cognitive, Affective, & Behavioral Neuroscience (2021) 21:156–171
https://doi.org/10.3758/s13415-020-00857-7

# Effects of trait empathy and expectation on the processing of observed actions

Christine Albrecht[1] · Christian Bellebaum[1]

Accepted: 23 November 2020 / Published online: 9 December 2020


## Abstract

Recent evidence suggests that the processing of observed actions may reflect an action prediction error, with more pronounced mediofrontal negative event-related potentials (ERPs) for unexpected actions. This evidence comes from an application of a false-belief task, where unexpected correct responses elicited high ERP amplitudes. An alternative interpretation is that the ERP component reflects vicarious error processing, as objectively correct responses were errors from the observed person's perspective. In this study, we aimed to disentangle the two possibilities by adding the factor task difficulty, which varied expectations without affecting the definition of (vicarious) errors, and to explore the role of empathy in action observation. We found that the relationship between empathy and event-related potentials (ERPs) mirrored the relationship between empathy and behavioral expectancy measures. Only in the easy task condition did higher empathy lead to stronger expectancy of correct responses in the true-belief and of errors in the false-belief condition. A compatible pattern was found for an early ERP component (150–200 ms) after the observed response, with a larger negativity for error than correct responses in the true-belief and the reverse pattern in the false-belief condition, but only in highly empathic participants. We conclude that empathy facilitates the formation of expectations regarding the actions of others. These expectations then modulate the processing of observed actions, as indicated by the ERPs in the present study.

**Keywords** Action observation · Expectation · Empathy · ACC · Error processing

Monitoring one's actions plays an important role in goal-directed behavior, making it possible to adapt performance quickly when necessary. An important aspect of this is the recognition of committed errors. For example, when you open the top drawer in the kitchen looking for a spoon, although you know that spoons are in the bottom drawer. In this case, you usually notice your error immediately. The neural processing of own errors has been thoroughly investigated over the past 30 years. In the 1990s, researchers first described a negative deflection in the event-related potentials (ERPs) of electroencephalography (EEG) data after error commission (Falkenstein, Hohnsbein, Hoormann & Blanke, 1991). This component, peaking within 100 ms after error commission, is called error negativity (Ne) or error-related negativity (ERN; Falkenstein et al. 1991; Falkenstein, Hoormann, Christ & Hohnsbein, 2000; Gehring, Goss, Coles, Meyer & Donchin, 1993; see also Gehring, Liu, Orr & Carp, 2012; Holroyd & Coles, 2002).

Error monitoring, however, is not limited to own errors. A negative deflection in ERPs similar to the ERN has been demonstrated for study participants observing others' errors. This deflection is referred to as observer error-related negativity (oERN; van Schie, Mars, Coles & Bekkering, 2004; see also Koban & Pourtois, 2014). As the ERN (Falkenstein et al., 2000; Dehaene, Posner & Tucker, 1994; Ridderinkhof, Ullsperger, Crone & Nieuwenhuis, 2004; Taylor, Stern & Gehring, 2007; see also Gehring et al., 2012), the oERN displays a frontocentral topography and is believed to originate from the anterior cingulate cortex (ACC) (Miltner, Brauer, Hecht, Trippe, & Coles, 2004; van Schie et al., 2004; see also Koban & Pourtois, 2014). Recent findings also indicate the involvement of other brain regions. Ullsperger, Danielmeier, & Jocham (2014) suggest the posterior medial frontal cortex (pMFC) as a generator of performance monitoring components, including the anterior and posterior midcingulate cortex, as well as presupplementary and

✉ Christine Albrecht
christine.albrecht@hhu.de

1 Institute of Experimental Psychology, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany

supplementary motor areas and the posterior dorsomedial prefrontal cortex. For action observation specifically, the superior temporal sulcus might additionally contribute to oERN generation (Ninomiya, Noritake, Ullsperger, & Isoda, 2018). In comparison to the ERN, the amplitude of the oERN is smaller (van Schie et al., 2004; Miltner et al., 2004). Not surprisingly, it also peaks later relative to the eliciting event (see Gehring et al., 2012), because it is not time-locked to a self-performed response but to an observed action. Moreover, its latency seems to vary between 130 and 300 ms, depending on the experimental paradigm (Bates, Patel, & Liddle, 2005; Koban, Pourtois, Vocat, & Vuilleumier, 2010 as opposed to Carp, Halenar, Quandt, Sklar & Compton, 2009; de Bruijn & von Rhein, 2012; Miltner et al., 2004; van Schie et al., 2004). The monitoring of others' actions can be considered a social process. For example, it is of particular importance for joint actions, when own actions are synchronized with others' (Loehr, Kourtis, Vesper, Sebanz, & Knoblich, 2013; Moreau, Candidi, Era, Tieri, Aglioti, 2020).

In recent years, the understanding of how performance monitoring is represented in the human brain and of the processes that underlie the ERN and related ERP components has changed. Increasing evidence supports the assumption that unexpected events, rather than errors, mainly drive ERP components and brain activity previously associated with error commission or error feedback for self-performed actions (Alexander & Brown, 2011; Ferdinand, Mecklinger, Kray, & Gehring, 2012; Wessel, Danielmeier, Morton, & Ullsperger, 2012). As accuracy and expectancy are usually confounded, at least for easy tasks, in which errors are rare, it cannot be differentiated whether an ERP component reflects an error or an unexpected event. This further applies to the ERPs associated with observed errors: Do they actually reflect the accuracy or the expectancy of these actions or both? It is assumed that when other people's actions are observed, predictions are formed that are then compared to the actually performed actions (i.e. the outcome of the prediction). If the two do not match, an action prediction error occurs (Burke, Tobler, Baddeley, & Schultz, 2010; Donnarumma, Costantini, Ambrosini, Friston & Pezzulo, 2017; Flanagan & Johansson, 2003), which is independent of the valence of the response, i.e., equally pronounced for an unexpected error and for an unexpected correct action.

An expectancy effect on a mediofrontal ERP component for observed actions has indeed been demonstrated in a previous study by our group. In that study, we applied a paradigm in which participants observed a person playing a stimulus-response task, the two-shell-game (see *Methods* for details). In this game, participants have to track under which of two shells a ball is hidden. Because this task is quite easy, erroneous responses by the observed person were unexpected (Kobza & Bellebaum, 2013). The task, however, also entails a false-belief condition: in this, observers had exclusive access to task-related knowledge that made correct responses unexpected. In this condition, the mediofrontal ERP component showed larger negative amplitudes after (unexpected) correct than (expected) error responses.

This finding appears to support the assumption that negative medio-frontal ERPs reflect that something unexpected happens. However, there may be an alternative interpretation. In the task we applied, as in real life, the observed person's subjective error could differ from the actual, objective error. To return to the introductory example: When you know that the spoons have been moved to the top drawer, but the observed person does not, then opening the top drawer looking for a spoon is objectively correct, but an error from the observed person's point of view. Objective and subjective error are dissociated in a false-belief condition, but not in the true-belief condition. Thus, the mediofrontal ERP component may also code vicarious error processing: Both conditions for which higher amplitudes were found, (objective) errors in the true-belief condition and (objectively) correct actions in the false-belief condition, are subjective errors to the naïve observed person. This interpretation in terms of vicarious error processing appears to be supported by a recent study where we found that trait empathy, measured by the empathy quotient (EQ) (Baron-Cohen & Wheelwright, 2004), was related to the processing of those actions in the two-shell-game that represented errors from the observed person's perspective (Bellebaum, Ghio, Wollmer, Weismüller, & Thoma, 2020). In participants with higher empathy scores, particularly large amplitudes of the mediofrontal negative ERP component were found in these conditions. This finding, however, was interpreted in terms of a facilitatory effect of empathy on the generation of expectations regarding observed actions. To summarize, it is not clear what cognitive process is primarily reflected in ERPs following observed actions—that is, whether they represent (subjective) accuracy from the perspective of the observed person and thus vicarious error processing or the (un)expectedness of the observed action, nor what role empathy plays in this respect. Although it shares some features with the oERN as described in the literature (van Schie et al., 2004), we will refer to the ERP component(s) of interest as negative mediofrontal ERP component in order to leave its functional significance undetermined.

In the present study, we aimed to disentangle effects of vicarious errors and action expectancy on the processing of observed actions by adding the factor task difficulty, because it should affect the latter but not the former. The two-shell game described above (Kobza & Bellebaum, 2013) can be considered an easy task (low level of difficulty), yielding clear expectations regarding the upcoming response in terms of accuracy, with correct responses being expected in the true-belief condition and errors in the false-belief condition. We reasoned that in a task of high difficulty, expectations would not be as clear. As there were only two response options, observers should expect that the observed person guesses more often, so that expectations concerning accuracy of the observed response would be nearer to chance level. In the introductory example, increasing task difficulty could correspond to looking for the drawer with the spoons in an unfamiliar kitchen, where you can more or less only rely on guessing. Comparing ERPs elicited by responses in high and low difficulty trials allows us to disentangle effects of action expectancy and vicarious error processing. The correct and incorrect answers remain the same for both high and low difficulty trials—subjectively, from the observed player's point of view, and also objectively. If the mediofrontal ERP component reflects vicarious error processing, no effect of task difficulty would thus be expected. However, if the observers' expectancy of the observed action determines its processing, task difficulty should have an effect. This notion is only true, however, if task difficulty indeed affects the expectancy of the observed response.

We hypothesized that the amplitude of the mediofrontal ERP component in response to observed actions that we and others described before (Bellebaum et al., 2020; Bates et al., 2005; Kobza & Bellebaum, 2013; van Schie et al., 2004) primarily reflects the expectancy of the observed responses rather than vicarious errors. In addition, we aimed to clarify the role of trait empathy in the processing of observed responses. By adding the factor task difficulty, we aimed to create more variance concerning the expected accuracy of the observed action, so that not only effects of expectancy and vicarious errors could be dissociated, but also the relationship between empathy and expectation effects regarding observed error monitoring could be examined.

# Methods

## Participants

A total of 38 participants took part in the study. As Mixed Linear Models are not yet used comprehensively and methods for power calculations have only emerged in the last years and require effect sizes for specific effects and interactions (Green & MacLeod, 2016), we chose this sample size based on studies using correlations to investigate the effect of continuous measures of trait empathy on action monitoring (Lockwood, Apps, Roiser, & Viding, 2015; Newman-Norlund, Ganesh, van Schie, de Bruijn, & Bekkering, 2009; Shane, Stevens, Harenski, & Kiehl, 2009). In these studies, sample sizes were between 20 and 31 participants. Five of the acquired participants were excluded from data analyses, either due to technical problems (four) or because the dependent variables derived from the EEG data were outliers in the analysis (one, see below for details). The remaining 33 participants (12 men) were between 18 and 33 years old ($M = 22.8$, $SD = 3.6$). They reported no previous or existing psychiatric or neurological illnesses and took no regular medication that could affect the nervous system. All participants had normal or corrected-to-normal vision and were German native speakers. Participants received course credit for taking part in the experiment. The study was approved by the ethics committee of the Faculty of Mathematics and Natural Sciences at Heinrich Heine University Düsseldorf, Germany.

## Experimental Task

The paradigm in this study was an adaptation of the two-shell game used by Kobza and Bellebaum (2013) and Bellebaum et al. (2020). Participants were asked to observe another person as he played the game. Unbeknownst to the participants, the player was fictitious and the displayed trials were simulated. The (fictitious) male player was introduced with a name and a photo, in order to give the impression that the participants observed the performance of a real person. The game started with a ball being hidden under one of two shells. After multiple rotations of the two shells (2, 3, or 4 rotations, randomly determined), the fictitious player pointed a joystick towards the shell where he believed the ball to be hidden. The observers saw the game from above, which also meant that they could see the ball at any time during the trial and therefore knew immediately whether the observed player was right or wrong when he moved the joystick at the end of the trial. The player's responses were balanced: half were correct responses (pointing to the shell covering the ball), and the other half were errors (pointing to the empty shell).

We aimed to modulate the observer's expectations concerning the player's responses by two factors. First, as in Kobza and Bellebaum (2013), a false-belief condition was introduced. That meant that the player was tricked in half of the trials (factor Trial Type): the ball was swapped between the two shells during one of the rotations. Observer participants were told that this was almost never visible to the player, while it was clearly visible to the observers themselves. If the player was tricked, the observers should expect a wrong rather than a correct answer of the player, because they believed that the player could not have seen the trick, and he would therefore assume that the ball was under the wrong shell. In the no-trick condition, respectively, the observers should expect the player to answer correctly.

As correct responses in the trick condition were errors from the perspective of the player, we added the factor Difficulty to the task, which aimed to disentangle vicarious errors and expectancy: the difficulty of keeping track of the ball was high in half of the trials, in that the shells were rotated more than three times faster (255 ms per rotation) than in the previous version of the experiment (850 ms per rotation), which was now considered the "slow" and thus low difficulty condition. Participants were told that due to the speed, it would be more difficult for the player to follow the shells with his eyes, so he had to guess more often in his decision under which shell the ball was. We assumed therefore that in low difficulty trials, observer participants would have stronger expectations regarding the player's response accuracy than in the high difficulty condition, for which the expected accuracy would only be slightly higher or lower than chance level (i.e., 50%) in the no-trick and trick conditions, respectively.

The experiment was arranged in four blocks of 117 trials between which participants could take short breaks. In contrast to the procedure in our previous studies applying this paradigm (Kobza and Bellebaum, 2013; Bellebaum et al., 2020), trick and no-trick conditions were alternated between blocks, because otherwise the build-up of expectations concerning response accuracy by the observed person might have been too complex given that we introduced an additional factor. In two of these blocks, the player was always tricked, in the other two blocks, he was never tricked. Observers knew in advance that the next block would only contain trick or no-trick trials. The order of the blocks was balanced between participants; either the first two blocks were trick-trials and the last two no-trick trials, or vice versa. Before the first block was started, participants completed 12 practice trials (6 trick and 6 no-trick trials).

Half of the 234 trials of each Trial Type were high difficulty trials, the other half were low difficulty trials. The low difficulty and high difficulty trials were presented in random order in the two blocks of each of the two levels of the Trial Type factor (trick and no-trick). In half of the trials, the fictitious player answered correctly, in the other half, he answered incorrectly, by pointing a joystick either at the shell containing the ball or at the empty shell (factor Accuracy). In total, there were thus eight conditions: correct and erroneous observed responses in low difficulty trick trials, high difficulty trick trials, low difficulty no-trick trials and high difficulty no-trick trials. It was pseudo-randomized on which side (left or right) the ball was located at the start and the end of each trial and how long the trial lasted (two, three or four rotations).

Twelve trials of each Trial Type and Difficulty did not end with the player's answer, but with the observer participants being asked which shell they thought the player would have chosen. After a static display of 400 ms of the final position of the shells, the respective question appeared ("Where will Daniel point the joystick?") as well as the letters "L" and "R" for left and right under the corresponding shells. Trials ended after button press or after 2,700 ms if no response had been given until then. These prompts aimed to provide an insight into the observer's expectations and were thus important to determine whether the intended manipulation of the observer participants' expectations worked.

A total of 420 trials were included in the EEG analysis, 105 trials for each combination of Trial Type and Difficulty. Forty-eight trials were included in the behavioral analysis: 12 of each Trial Type and Difficulty condition. The time course of the individual trials is shown in Figure 1.

## Empathy measure

Participants were asked to complete the German version of the Cambridge Behavior Scale (Baron-Cohen & Wheelwright, 2004; de Haen, n.d.), which is a measure of trait empathy. In a previous study (Bellebaum et al., 2020), we found that this empathy measure interacts with the experimental factors Trial Type and Accuracy of the paradigm that we (in an adapted version that additionally includes Difficulty) also applied in the present study, which is why we focused on this measure. The questionnaire contains 60 items, 20 of which are distractor items. Items consist of statements (e.g., "I really enjoy caring for other people"), which the participants can agree or disagree with using a four-point Likert scale ranging from "strongly agree" to "strongly disagree." Items are scaled negatively or positively. Participants can score a maximum of two points per item. For positively scaled items, participants receive two points if they "strongly agree," one point if they "slightly agree" and zero points if they "slightly disagree" or "strongly disagree." For negatively scaled items, the scoring is reversed. Participants do not receive points for any answer on distractor items. Points are added and result in an empathy quotient (EQ) sum score that can range from 0 to 80.

## EEG Recording

Thirty passive scalp electrodes were applied according to the international 10-20 system (F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, CP3, CPz, CP4, P7, P3, Pz, P4, P8, PO7, PO3, POz, PO4, PO8), and an electroencephalogram (EEG) was recorded throughout the experiment using a BrainAmp Standard amplifier (Brain Products, Munich, Germany) and the corresponding software (BrainVision Recorder, version 1.20.0506, Brain Products, Munich, Germany) at a sampling rate of 1,000 Hz. Electrodes were referenced to the average of two electrodes on the left and right mastoids. All impedances were kept below 5 kΩ.

## Procedure

Upon arrival in the laboratory, participants were informed about the experimental procedure and gave written informed consent to participate in the study. They were then asked to fill in a demographic questionnaire and the German version of the Cambridge Behavior Scale (Baron-Cohen & Wheelwright, 2004; de Haen, n.d.). After completion, we attached the EEG electrodes and participants were placed in front of a 1,920 * 1,080 px desktop monitor, and the experiment began. The experiment lasted about 45 minutes. The Stimulus

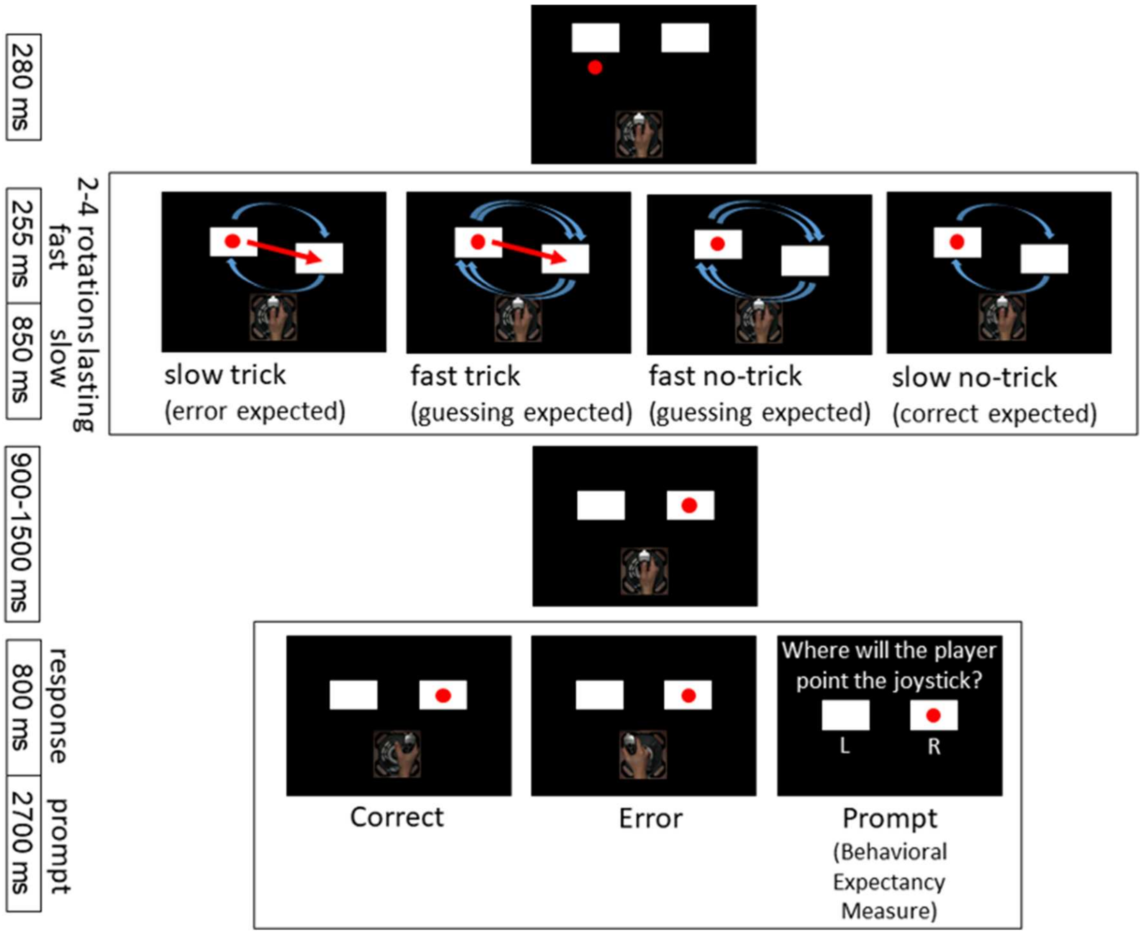



**Figure 1.** Time course of events in the experiment trials. There were eight conditions, low difficulty no-trick, low difficulty trick, high difficulty no-trick, and high difficulty trick, which either ended in a correct or an error response. Some trials ended not in a response by the observed player but in a prompt question to measure the observer's expectancies.

presentation and response recording were controlled by Presentation Software (Version 20.0, Neurobehavioral Systems, Albany, CA).

## Data analyses

**Behavioral data** We analyzed the responses to the prompt trials to determine how the observers' expectancies concerning response accuracy of the player were modulated by the factors Trial Type and Difficulty. As in our previous studies (Kobza and Bellebaum, 2013; Bellebaum et al., 2020), we aimed to induce expectations of correct responses in no-trick trials and of error responses in trick trials, which were possibly less strong in high difficulty trials. We thus determined the proportion of the prompt trials in which the observer participants expected a correct response by the player for low difficulty and high difficulty no-trick and trick trials.

**EEG data** EEG data were preprocessed using BrainVision Analyzer software, version 2.1 (Brain Products, Munich, Germany). Raw data were filtered with a 0.5-Hz high-pass and a 20-Hz low-pass filter. We then aimed to remove blink artefacts from the filtered raw signal. For this purpose, we performed an independent component analysis on single-subject EEG data. This analysis breaks down the raw data into temporally independent and spatially fixed components. We selected one component per participant that seemed to represent blink and vertical eye movement artifacts as observed in the vEOG electrode, as indicated by a symmetrical frontal distribution across the scalp. This component was then removed via independent component analysis back-transformation. For ERP analysis, we created segments of 800-ms length that started 200 ms before the observed choice (the time point when the joystick pointed to one of the shells). We performed a baseline correction using the average signal in the 200 ms before the observed choice. All segments in which a voltage step larger than 50 μV per ms occurred, in which highest and lowest data points differed by more than 100 μV or in which signals at any sample were higher than 100 μV or lower than −100 μV were excluded from further analysis automatically. On average, 3% of the error no-trick high difficulty segments (SD = 7%), 3% of the correct no-trick high difficulty segments

(SD = 7%), 4% of the error no-trick low difficulty segments (SD = 8%), 4% of the correct no-trick low difficulty segments (SD = 7%), 2% of the error trick high difficulty segments (SD = 5%), 3% of the correct trick high difficulty segments (SD = 4%), 3% of the error trick low difficulty segments (SD = 4%), and 3% of the correct trick low difficulty segments (SD = 4%) were excluded. None of the participants lost more than 30% of all segments. Finally, single-subject averages were created for all eight conditions of the experiment. Data for each subject and condition were exported as text files and further processed in MATLAB, version R2017b (Mathworks, Natick, MA).

Based on the findings obtained in previous studies of our research group using this paradigm (Kobza & Bellebaum, 2013; Bellebaum et al., 2020), we expected that our experimental manipulation would affect a negative-going component in the ERPs between 250 and 420 ms after the observed choice. For this component, an interaction between the factors Trial Type and Accuracy (Kobza & Bellebaum, 2013; Bellebaum et al., 2020) as well as a modulation of this interaction by Empathy (Bellebaum et al., 2020) have been described. Thus, we investigated this component first. As in our previous studies, we calculated a peak-to-peak amplitude for the negative peak relative to a preceding positive peak. First, we pooled the signal over the electrodes Fz, FCz, and Cz, at which the component was particularly pronounced (see Bellebaum et al., 2020, for a similar procedure). We then calculated the most negative peak between 250 and 420 ms after the observed choice and subtracted the most positive peak in the preceding time window between 130 ms and the negative peak.

Contrary to our hypotheses, the ERPs seemed to be modulated by the experimental manipulations also in an earlier time window. Visual inspection of the signal at frontocentral electrodes suggested that the experimental factors modulated the first negative peak, that is, the N1 amplitude. Such an early modulation was not entirely unexpected: for the oERN, for example, as an ERP component reflecting the processing of observed actions, some variability has been found in studies concerning the latency with which it occurs. While it has mostly been reported to peak later than 200 ms after the response (van Schie et al., 2004; Miltner et al. 2004; Kobza & Bellebaum, 2013), there also are reports of shorter latencies (see Koban & Pourtois, 2014), with peaks as early as 150 ms after the onset of the observed response in some studies (Bates et al., 2005). It thus seems conceivable that a modulation of the processing of observed actions can take place in the N1 time window. To analyze this component, we also considered the pooled signal of three frontocentral electrodes (Fz, FCz, and Cz), because the component was also most pronounced frontocentrally (see topographic maps in the *Results* section). To score the component, we determined the most negative peak between 100 and 250 ms and subtracted the preceding most positive peak between 50 ms and the negative peak of the pooled signal. We will refer to this component as the early frontocentral negative component, while the component we analyzed first (see Bellebaum et al., 2020) will be referred to as the late frontocentral negative component.

**Outlier detection** In each of the two EEG data sets (early component, late component) we determined participants whose peak-to-peak amplitude in at least one of the eight conditions differed by more than three standard deviations from the mean to identify outlier values in these dependent variables. The same criterion was used for the EQ sum score as continuous predictor variable. We found one participant with elevated scores for both dependent variables and excluded this participant from further data analysis.

**Statistical analysis** The statistical analyses of the behavioral and EEG data were based on the following strategy. First, the behavioral data were analyzed to show if the task Difficulty factor, together with the Trial Type factor, affected observers' expectancy in the intended way. Specifically, an interaction between the Trial Type and Difficulty factors was expected, with pronounced expectations concerning the accuracy of the observed action emerging only in low difficulty trials. In a second step, we analyzed to what extent Difficulty and Trial Type (due to their hypothesized effect on expectancy) affected the processing of observed actions, as reflected in the early and late frontocentral negative ERP components. If Trial Type and Difficulty interact in their effect on expectancy, the two factors also should interact in their effect on an ERP component reflecting expectancy. The focus in the analysis was thus on interactions involving these two factors. This analysis procedure established an indirect link between the behavioral (expectancy) data and the neurophysiological data. In addition, potential effects of the continuous factor Empathy were considered. We analyzed our data by means of linear mixed effects (LME) analyses using the lme4 statistical package (version 1.1-21) in R (version 3.5.3), because this type of analysis allows to include both categorical and continuous factors. All models were estimated using a restricted maximum likelihood approach, as proposed by Luke (2017). The R package lmerTest (version 3.1-0) was applied for evaluating significance in the models by using Satterthwaite approximation for the degrees of freedom. Effect sizes were calculated with the function anova_stats of the package sjstats (version 0.17.9).

For the behavioral data, we defined the dependent variable as the percentage of prompt trials in which participants indicated that they expected the player to choose the correct answer. We thus specified a model with percentage of expected correct answers as dependent variable and participants as a random-effect factor. Trial Type and Difficulty were defined as categorical fixed-effect predictors. We also included the random slopes of the categorical predictors by participants.

As continuous factor we included Empathy (the EQ sum score). For the categorical factors, the levels of the Trial Type factor were recoded as +1 for trick and −1 for no-trick and for the factor Difficulty as +1 for low difficulty and −1 for high difficulty. The Empathy measures were mean-centered.

In the subsequent analysis of the EEG-data we analyzed the later as well as the earlier frontocentral negative ERP component. We thus specified two models, one for each of the components as dependent variable, that were similar to the model for the behavioral data, with Trial Type and Difficulty and the additional factor Accuracy as categorical fixed-effect predictors (modelling also their random slope by participants). Trial Type and Difficulty were recoded as in the model for the behavioral data (+1 = trick; −1 = no-trick; +1 = low difficulty; −1 = high difficulty). Accuracy was recoded as +1 for correct and −1 for error responses. The continuous factor, Empathy, and the model estimation were the same as in the model for the behavioral data. For all analyses the threshold for statistical significance was set to $p < .05$.

*Interactions*. Interactions were resolved in a step-wise manner according to Aiken, West, & Reno (1991): for every n-way interaction, we calculated slopes of the n-1-way interactions while one predictor was held constant. Significant interactions in these analyses were then resolved in the same way until all factors were resolved. For categorical factors, in accordance with the variable coding, we used 1 or −1 as constants. For the continuous factor Empathy, we shifted the variable by one standard deviation downward or upward from the mean ($M$ – 1 $SD$ or $M$ + 1 $SD$) and calculated lower-level interactions while holding the continuous factor constant at low level empathy (low empathy, $M$ – 1 $SD$) or high level empathy (high empathy, $M$ + 1 $SD$).

*Analysis linking behavioral and ERP measures.* To explore whether there also was a direct relationship between expectancy and observed response processing, we planned to conduct follow-up analyses in case of a significant effect of the Trial Type and Difficulty factors on one of the ERP components. For this purpose, we calculated expectancy measures (concerning correct responses) for each of the conditions trick high difficulty, trick low difficulty, no-trick high difficulty, and no-trick low difficulty in every participant based on the prompt trials. These values indicated how strongly correct responses were expected. For error responses, the expectancy values were recalculated as 1 – expectancy of the correct response. The expectancy values were used as continuous factor Expectancy (mean-centered) in an LME model, including ERP component amplitudes as dependent variable. We included all participants that were included in the previous analyses in an additional outlier detection, based on the so-called Cook's Distance. As Cook's Distance measures the influence of single subjects on the model, this outlier detection method might be especially suitable for exploratory analyses where some aspects might not be perfectly controlled for (e.g., correlations between the predictors). Cook's Distance analysis revealed one outlier participant that was excluded from the Expectancy analysis. To further test whether Empathy explained additional variance beyond the effect of expectancy, we calculated an exploratory Chi-Square test using the anova-function in R (from the package car, v 3.0-9) to compare the two models. This allowed us to determine whether a model including Expectancy and Empathy explained significantly more variance than a model including only Expectancy and thus, whether the frontocentral negative ERP component is further influenced by trait Empathy. For this comparison, both models were recalculated with a maximum likelihood approach.

## Results

Please find additional statistical data for the following LME analyses in the Supplementary Materials, including estimates, t-test statistics, standard errors, and confidence intervals for the data in the reported analyses.

### Behavioral analysis

The behavioral data reflecting the strength of the expectations of the observers are depicted in Figure 2. For the percentage of expected correct responses we found significant effects of Trial Type, $F(1,31.00) = 25.94$, $p < 0.001$, $\eta_p^2 = 0.23$, and Difficulty, $F(1,31.00) = 11.55$, $p = 0.002$, $\eta_p^2 = 0.12$. Trick trials resulted in lower expectation of correct answers ($b = −12.41$) than no-trick trials, as did high difficulty trials ($b = 8.24$) compared with low difficulty trials. Furthermore, a significant interaction between Trial Type and Difficulty emerged, $F(1,30.99) = 8.85$, $p = 0.006$, $\eta_p^2 = 0.09$. As can be seen in Figure 3, expectancy of correct responses was descriptively nearer to chance level in the high difficulty than low difficulty trials. A follow-up analysis to resolve the interaction of the two factors revealed, however, that the factor Difficulty was only significant for no-trick trials, $F(1,55.91) = 20.18$, $p < 0.001$, $b = 13.35$, not for trick trials ($p = 0.299$). In no-trick trials, correct answers were more expected in low difficulty trials than in high difficulty trials. Analyzing high and low difficulty trials separately, we found that Trial Type was significant for both low difficulty trials, $F(1,55.75) = 35.63$, $p < 0.001$, and high difficulty trials, $F(1,55.75) = 5.98$, $p = 0.018$, but the difference was less pronounced in high difficulty trials ($b = −7.30$ as opposed to $b = −17.53$). In both Difficulty conditions, correct answers were more expected in no-trick trials. As Difficulty interacted with Trial Type and the difference between trick and no-trick trials was less pronounced in high difficulty trials, we can conclude that the difficulty manipulation worked, as expected.

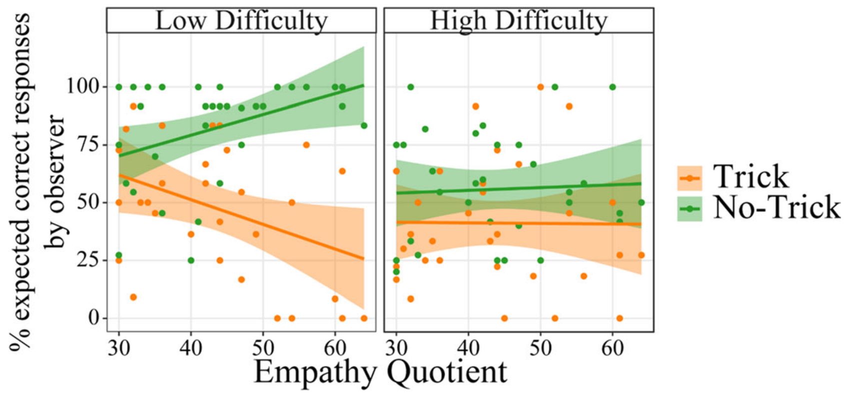



**Figure 2.** Behavioral data derived from prompt trials. Displayed is the percent of trials in which participants stated that they expected the player to answer correctly, modulated by trait Empathy, Trial Type, and trial Difficulty. Confidence intervals are displayed around the regression lines.

Significant differences between low and high difficulty trials were found, however, only for no-trick, but not for trick trials.

We also found that Empathy interacted with Trial Type, $F(1,31.00) = 4.59$, $p = 0.040$, $\eta_p^2 = 0.05$, and with Trial Type and Difficulty, $F(1,30.99) = 6.87$, $p = 0.013$, $\eta_p^2 = 0.07$. We further investigated the interaction effects, including the continuous factor Empathy. We first considered the interaction between Empathy and Trial Type. Although the factor Trial Type modulated the answer for participants with both high and low empathy, $F(1,31.00) = 26.14$, $p < 0.001$, and $F(1,31.00) = 4.32$, $p = 0.046$, respectively, the effect was larger for high empathy participants ($b = -17.66$) than for low empathy participants ($b = -7.17$). For both groups, trick trials resulted in a lower expectation of correct answers than no-trick trials, but the effect was larger in participants with high empathy. Subsequently, we resolved the three-way interaction. A Trial Type by Empathy interaction was significant only for low difficulty trials, $F(1,55.75) = 10.64$, $p = 0.002$, not for high difficulty trials ($p = 0.812$). Further simple-slope analyses for low difficulty trials revealed that the factor Empathy was significant for low difficulty trick trials, $F(1,40.83) = 5.31$, $p = 0.026$, as well as low difficulty no-trick trials, $F(1,45.66) = 5.34$, $p = 0.025$. Higher empathy led to lower expectation of correct answers in low difficulty trick trials ($b = -1.07$) and to higher expectation of correct answers in low difficulty no-trick trials ($b = 0.90$). The main effect of Empathy and the remaining interaction did not reach significance (all $p \geq 0.786$).

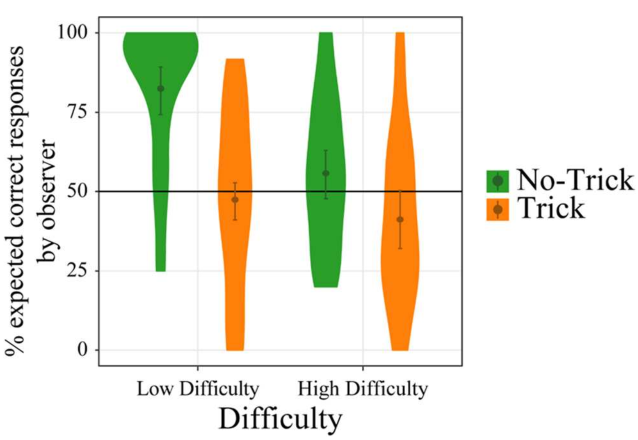



**Figure 3.** Interaction effect between Trial Type and trial Difficulty in behavioral data derived from prompt trials. The black line marks chance level (50%). Mean and confidence intervals are displayed within the respective violin plots.

In summary, we found that the expectancy modulation by the factors Trial Type and Difficulty succeeded (Figure 3). Importantly, expectations were further modulated by empathy, with higher effects of the experimental factors on expectancy measures in high empathy participants. However, even for high empathy participants, we only found a significant effect of Difficulty in no-trick, but not in trick trials (Figure 2).

## EEG analysis

**Late frontocentral negative component** The ERPs and their topography for the relationship between Trial Type, Difficulty and Accuracy are depicted in Figure 4. A display of the relationship between the four factors and the amplitude of the late frontocentral negative component, whose mean latency (across participants and conditions) was 335 ms (*SD* = 52 ms), is shown in Figure 5. The LME analysis did not reveal any main effects for the late negative component amplitude (all $p \geq 0.149$). We found one significant interaction between Difficulty and Accuracy, $F(1,124.84) = 11.77$, $p < 0.001$ (for a visualization of this effect, see Figure 6). Follow-up analyses revealed that Accuracy modulated the late negative component only in high difficulty trials, $F(1,59.99) = 4.75$, $p = 0.033$, $\eta_p^2 = 0.06$, not in low difficulty trials ($p = 0.130$). For high difficulty trials, errors elicited a larger amplitude ($b = 0.42$) than correct responses. We did not find any other interaction effects (all $p \geq 0.323$).

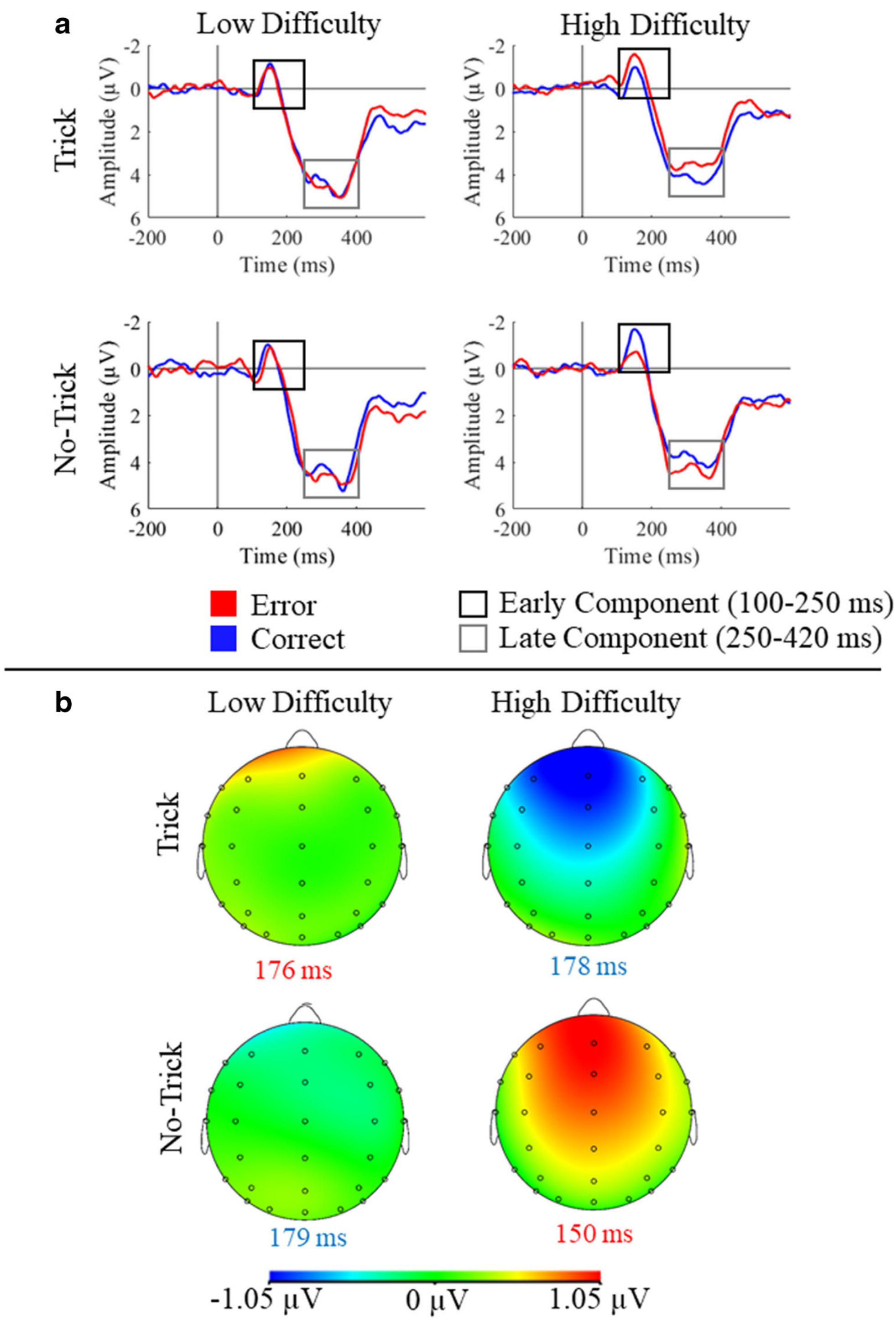



**Figure 4.** **A**. ERPs pooled over Fz, FCz, and Cz after observed correct and error responses for all combinations of the Trial Type and Difficulty conditions. The two analyzed components, early and late frontocentral negative component, are marked in the ERPs. **B**. Topography of the difference between the ERPs after error and correct responses at the maximum positive (low difficulty trick and high difficulty no trick) or the maximum negative (low difficulty no trick and high difficulty trick) peak of the difference between error and correct responses (between 150 and 180 ms) for the pooled signal of Fz, FCz, and Cz for both Trial Type and Difficulty conditions.

In summary, we did not find the expected modulation of ERPs in the late frontocentral negative ERP component (Figure 5), as the result pattern did not mirror the expectancy modulation by the factors Trial Type and Difficulty in the form of an interaction between the factors. Instead we found a selective modulation of the late component by Accuracy in high difficulty trials (Figure 6).

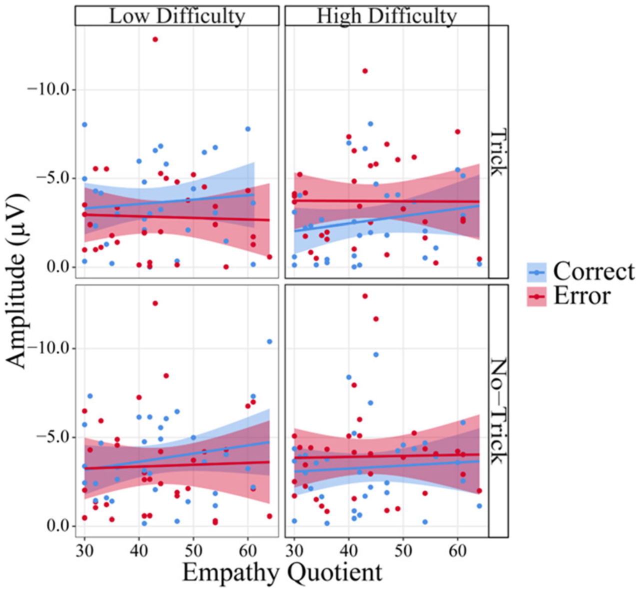



**Figure 5.** Peak-to-peak amplitudes of the late frontocentral negative component (250–420 ms) as a function of Trial Type, Difficulty, Accuracy and Empathy. Confidence intervals are displayed around the regression lines.

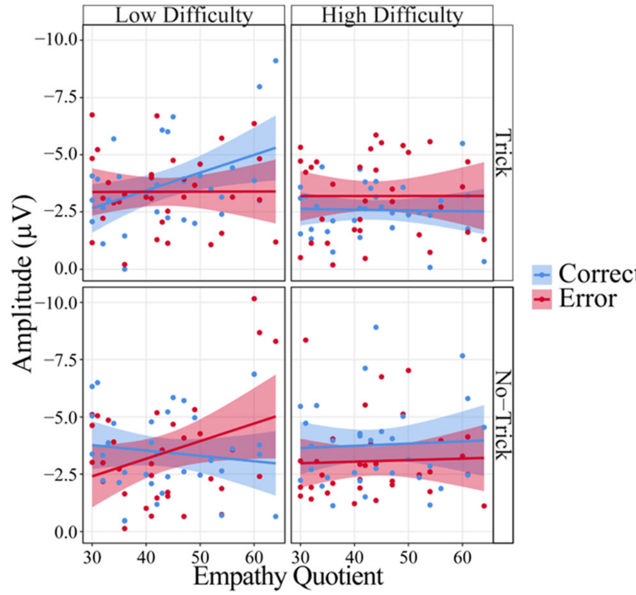



**Figure 7.** Peak-to-peak amplitudes of the early frontocentral negative component as a function of Trial Type, Difficulty, Accuracy, and Empathy. Confidence intervals are displayed around the regression lines.

**Early frontocentral negative component** Figure 7 shows the relation between the four factors and the amplitude of the early frontocentral negative component. The mean latency of this component (across participants and conditions) was 159 ms ($SD$ = 27 ms). The LME analysis on this component's amplitude revealed no significant main effects and no two-way interactions (all $p \geq 0.069$). Instead we found two three-way interactions, one between Empathy, Trial Type and Accuracy, $F(1, 186.02) = 6.04$, $p = 0.015$, $\eta_p^2 = 0.03$, and one between Trial Type, Difficulty and Accuracy, $F(1{,}186.02) = 5.64$, $p = 0.019$, $\eta_p^2 = 0.03$, but no other three-way-interactions (all $p \geq 0.524$). Because we also found a significant four-way interaction between all four factors—Empathy, Trial Type, Difficulty, and Accuracy, $F(1, 186.02) = 6.96$, $p = 0.009$, $\eta_p^2 = 0.03$, we focused on the resolution of this highest-order interaction. We thus conducted two further LME-analyses separately: one for low difficulty and one for high difficulty trials. A significant three-way interaction between Trial Type, Accuracy, and Empathy emerged for low difficulty trials, $F(1{,}186.01) = 12.98$, $p < 0.001$, but not for high difficulty trials ($p = 0.898$). In the resolution of the interaction for low difficulty trials, an Accuracy by Empathy interaction was found for both low difficulty trick, $F(1{,}194.18) = 4.83$, $p = 0.029$, and low difficulty no-trick trials, $F(1{,}194.08) = 8.20$, $p = 0.005$. For low difficulty trick trials, a significant effect of empathy was found only for correct, $F(1{,}77.83) = 7.02$, $p = 0.010$, but not for error responses ($p = 0.984$); for no-trick trials the pattern was reversed: an effect of Empathy was found for error, $F(1{,}52.71) = 4.56$, $p = 0.037$, but not for correct trials ($p = 0.492$). Higher empathy resulted in more negative amplitudes for low difficulty correct trick trials ($b = -0.08$) and for low difficulty error no-trick trials ($b = -0.08$). Analyzing the high difficulty trials separately, no three-way interaction emerged ($p = 0.898$), but an interaction between Trial Type and Accuracy could be found, $F(1, 186.01) = 7.23$, $p = 0.008$. A significant effect of Accuracy emerged only for high difficulty no-trick trials, $F(1{,}194.20) = 4.10$, $p = 0.044$, where correct responses led to higher amplitudes ($b = -0.35$), although we found a trend for an Accuracy effect also in high difficulty trick trials, $F(1{,}194.16) = 3.07$, $p = 0.081$, where errors led to higher amplitudes ($b = 0.31$).

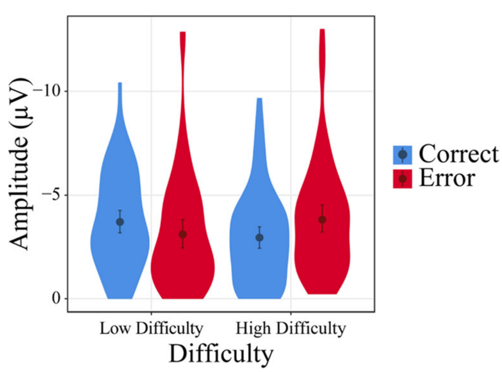



**Figure 6.** Interaction effect between Difficulty and Accuracy for the late frontocentral negative component (250–420 ms). Mean and confidence intervals are displayed within the respective violin plots.

In summary, we found a modulation in an earlier time window (around the N1) similar to the one we expected. Consistent with the behavioral results for expectancy formation, we found that ERPs were modulated by an interaction of Empathy, Trial Type, and Accuracy for low difficulty trials only, where Empathy seemed to be important particularly for the processing of those events that were considered unexpected (see the low difficulty grids in Figure 7).

### Effects of expectancy and empathy on the early frontocentral negative component

After we found that Trial Type and Difficulty affected the early frontocentral negative ERP component, we aimed to explore the relationship between expectancy and the amplitude directly (see *Methods* section). We found a main effect of Expectancy, $F(1,220.43) = 4.57$, $p = 0.034$, $\eta_p^2 = 0.02$, $b = 0.65$. For higher Expectancy values, amplitudes were smaller (Figure 8). We further conducted an analysis in which we compared a model including only Expectancy to a model including Expectancy and Empathy. This analysis must be considered exploratory, because empathy predicted expectancy in our experiment, as was revealed by the behavioral data analysis. The two factors in the Expectancy x Empathy model were thus not independent. A model including both predictors did not explain significantly more variance than a model including only Expectancy, $X^2(2) = 0.07$, $p = 0.967$. In summary, the measured Expectancy values functioned as predictors for the amplitude of the early frontocentral negative component. Empathy did not account for significantly more variance if Expectancy was already included as a predictor.

## Discussion

In this study, we investigated the role of expectations in the processing of observed actions and a potentially mediating effect of empathy. To this end, we had participants observe a person perform correct or error responses in a two-shell-game. Expectancy was modulated by two factors that allowed to distinguish between effects of vicarious errors and expectancy. We found that our manipulation of the expectancy of the observed response succeeded. We also found an effect of empathy on the strength of the expectations. Concerning the neurophysiological processing of observed responses, there was evidence that the amplitude of a frontocentral negative ERP component time-locked to observed responses was mainly driven by the expectancy of the responses. Surprisingly, this pattern was found in the N1 time window and thus earlier than the ERP components that have been linked to observed response processing in previous studies (Kobza and Bellebaum, 2013; Bellebaum et al., 2020; van Schie et al., 2004; see also Koban & Pourtois, 2014).

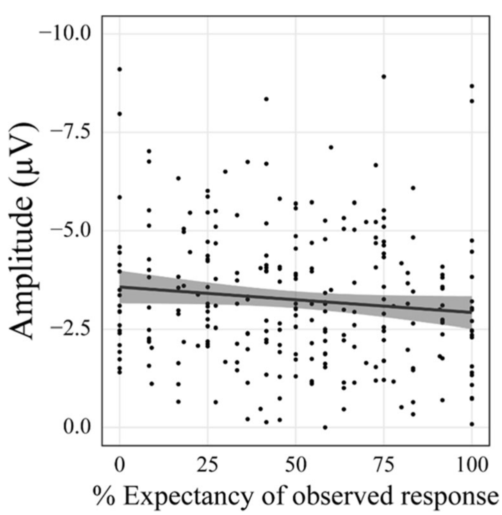



**Figure 8.** Effect of Expectancy on the early frontocentral negative component. Confidence intervals are displayed around the regression lines.

### Behavioral measures of expectancy

We measured self-reported expectancies concerning the observed response separately for each condition. This assessment served to verify whether our experimental manipulations affected participants' expectancies in the intended way, which was an important prerequisite for the interpretation of the ERP data. We found a modulation by a false-belief condition (factor Trial Type), in accordance with previous studies applying the same paradigm (Bellebaum et al., 2020; Kobza and Bellebaum, 2013). Moreover, the newly introduced factor task Difficulty affected participants' expectations in that the difference between conditions with true and false belief was less pronounced for trials with high difficulty. Analyzing true- and false-belief conditions separately, we found a modulation by task Difficulty only in the true-belief, but not in the false-belief condition. One reason seems to be that expectations in low difficulty trials with a false-belief were less strong than expectations for low difficulty true-belief trials. Expectations for false-belief conditions seem to be harder to form (the same effect was found in previous studies employing this paradigm, see Bellebaum et al., 2020; Kobza & Bellebaum, 2013; as well as in studies using false-belief tasks, see Birch & Bloom, 2007; Wellman, Cross, & Watson, 2001), and with the additional factor Difficulty, this might have led to smaller differences between the Difficulty conditions. We also found that Empathy influenced expectancy formation in low difficulty conditions: expectancies were formed most consistently in high empathy participants. Bellebaum et al. (2020) did not find a modulation of expectancy by empathy using the same paradigm. In this previous study there was little interindividual variance in the expectancies, which clearly differed between false- and true-belief conditions. The introduction of the task difficulty variation in the present study may have led to more variance, so that empathy may have plaid a more important role for expectancy generation.

It is important to keep in mind that expectancy was assessed based on 12 prompt trials per condition only, so the derived expectancy values do not reflect expectations on a single trial basis. Also, it has to be noted that the prompt trials primarily served to check if the expectancy manipulation succeeded. While the LME analysis suggests an interesting modulation of expectancy by empathy, the behavioral results should be interpreted with caution.

### Latency of expectancy and empathy effects on the processing of observed actions

An important difference between the present study and our previous studies with the same paradigm (Kobza and Bellebaum, 2013; Bellebaum et al., 2020) is that the modulation of the ERPs by Expectancy and Empathy occurred much earlier after the observed response in the present study. The component also had a lower latency than the oERN, at least for most of the studies investigating this component (Carp et al., 2009; de Bruijn & von Rhein, 2012; Miltner et al., 2004; van Schie et al., 2004); it occurred between 100 and 250 ms (mean 159 ms) and thus in the latency range of the N100. Nevertheless, we have reason to believe that this early modulation reflects an ERP component resembling other components that have been linked to the processing of observed responses. Firstly, the topography of the relative negativity in the ERPs for unexpected events showed a fronto-central maximum (Figure 4) and is thus not only consistent with the typical topography of the oERN, but also with that of the ERN and the feedback-related negativity, which are all related to performance monitoring (Falkenstein et al., 1991; Gehring et al., 2012; Gehring & Willoughby, 2002; Miltner et al., 2004; van Schie et al., 2004). Second, the modulation by expectancy and/or empathy is comparable to modulations of monitoring-related ERPs found in previous studies (Bellebaum et al., 2020; Ferdinand et al., 2012; Kobza & Bellebaum, 2013; Wessel et al., 2012). Particularly, these results correspond to those of Kobza and Bellebaum (2013) and Bellebaum et al. (2020), who applied the same paradigm but found the effect in a later time window. Third, and most important, the latency of the oERN and related components appears to differ depending on the task. For a Go/NoGo-Task an oERN was observed as early as 150 ms after the observation of errors in NoGo trials which corresponds to the latency range of the present study (Bates et al., 2005; Koban et al., 2010). This earlier latency has been linked to the lower complexity compared with a Flanker task (Koban & Pourtois, 2014).

The question remains, however, why the modulation in the present study occurred so early. The main difference between the present and our previous studies (Bellebaum et al., 2020; Kobza & Bellebaum, 2013) is that we used a block design for trick and no-trick trials, so that participants knew beforehand whether the observed person had a true or a false belief. In blocks with trick trials, for example, they knew that the observed person was more likely to point to the empty shell, performing an error. If trick and no-trick trials alternate trial-by-trial, as in our previous studies, expectation formation probably takes more time. We believe that this early expectation formation enabled faster processing and thus earlier ERP modulations. We therefore discuss the early ERP modulation in the following sections and relate it to findings from the literature, where mostly later components were analyzed, but emphasize that these results should be interpreted with caution as our hypothesis was related to a modulation in a later time window.

### Effects of expectation on observed error processing

Consistent with the previous results we obtained with this paradigm (Kobza and Bellebaum, 2013; Bellebaum et al., 2020), we found that the amplitude of a negative ERP component following an observed response was modulated by expectancy, although this modulation occurred earlier than in the previous studies In participants who developed strong expectations, the least expected events, that is, correct responses in low difficulty trials with a false-belief and error responses in low difficulty trials with a true belief, elicited the highest amplitudes.

Importantly, we found a modulation not only by the false-belief condition, but also by the new factor task Difficulty. If the ERP component had only reflected the false-belief condition and not task difficulty, the modulation could have been ascribed to vicarious error processing, as errors from the perspective of the observed person were the same in both Difficulty conditions. As this is not the case, we conclude that the ERP modulation codes expectancy rather than vicarious errors, which corresponds with the interpretation of Kobza and Bellebaum (2013) and Bellebaum et al. (2020).

This supports theories on the role of expectancy for amplitudes of ERP components generated by the ACC or pMFC, stating that these components primarily code unexpected events irrespective of valence (Alexander & Brown, 2011). It also matches other recent results. Wessel et al. (2012) found a common neural generator, namely the pMFC, of both the ERN and the novelty-related frontocentral N2, suggesting an overlap of the neural correlates of error and surprise processing. Ferdinand et al. (2012) showed that the FRN was elicited not only by unexpected negative, but also by unexpected positive feedback. For observed actions, Schiffer, Krause, & Schubotz (2014) reported activity in the medial prefrontal cortex after unexpected incorrect as well as unexpected correct responses in a functional magnetic resonance imaging study (see also Wang et al., 2015). Our study thus adds to existing evidence that activity in the medial prefrontal cortex and ACC in response to different events is critically modulated by the expectancy of these events.

It has to be noted that, in contrast to our hypotheses, the ERP amplitude pattern was reversed in high difficulty trials compared with low difficulty trials. Amplitudes were enhanced for correct actions in true-belief and erroneous actions in false-belief trials, respectively, which is not in accordance with the behavioral expectancy measures. We suspect that although participants formed explicit expectations, at least in the true-belief condition, further implicit expectations might have played a role, too. Multiple studies suggest that responses to errors or infrequent events lead to increased attention to the source of the (prediction) error (Notebaert, Houtman, Opstal, Gevers, Fias, & Verguts, 2009; Steinhauser & Andersen, 2019; Wessel, 2018). In our study, trick and no-trick trials were presented in separate blocks, while low difficulty and high difficulty trials were mixed within one block. This means that for every trial, the observer participants only had to consider whether the trial was difficult (high difficulty) or not (low difficulty) for their expectancy of the correct response, so that high difficulty and low difficulty trials were directly compared to each other. As participants had to focus on differentiating high and low difficulty trials, their attention might have been relocated to this comparison as opposed to absolute probabilities. The modulations observed might thus code for the comparison between high and low difficult trials, meaning that the likelihood of the actually more expected responses might be implicitly underestimated in high difficulty trials, resulting in the observed reverse pattern.

The exact mechanisms of expectancy formation, especially concerning explicit and implicit expectancies that might have played a role in our study, are not completely understood. In a study in which participants observed erroneous everyday actions in a virtual reality setting, Pezzetta, Nicolardi, Tidoni, & Aglioti (2018) found a modulation of the ERPs by the accuracy of the observed response, with higher oERN amplitudes for errors, also when errors occurred more frequently than correct actions. A reason for this could be that the authors used simple everyday actions which might generally be expected to be performed correctly. Furthermore, other studies suggest that the way events are processed is not entirely determined by their frequency. Several studies found differences between the processing of negative and positive feedback processing even if these events were equally probable (Wang et al., 2015; Yeung et al., 2005). These findings have been ascribed to an overoptimistic bias of participants to expect correct responses or positive outcomes more strongly, especially for own behavior (Oliveira, McDonald, & Goodman, 2007). Ferdinand et al. (2012) found comparable amplitudes for unexpected positive and negative feedback in the FRN, but observed an effect of valence in the P300, with positive feedback eliciting larger amplitudes than negative feedback. This difference in early and later processing resembles the difference between the early and the late frontocentral negative component in high difficulty trials in this study, with the early component being modulated by expectancy and the late component being modulated by valence. Also, Ferdinand et al. describe a difference between actual expectations (more than half of the participants believed that negative feedback was more frequent when asked after the experiment, only less than a quarter thought that positive feedback was more frequent) and FRN amplitudes, again suggesting that other, less conscious processes play a role in early processing when outcomes are uncertain. Moreover, other factors apart from expectancy may also affect neural indices of performance monitoring: Maier and Steinhauser (2016) found that active responders' ERN was modulated by error significance rather than error probability.

### Effects of empathy on observed error processing

Another similarity between the present study and our previous work on action observation is that empathy affected the processing of observed responses (Bellebaum et al., 2020). In the present study, this effect was restricted to the low difficulty condition, where only for highly empathic participants ERP amplitudes were higher for unexpected than expected events. Figure 7 shows that the processing of unexpected events (correct responses in low difficulty trick trials and error responses in low difficulty no-trick trials) was exclusively modulated by empathy. As outlined in the introduction, our previous finding might have reflected vicarious error processing, as in the false-belief condition of our task correct responses were errors from the observed person's view. However, by showing that the empathy effect is restricted to low difficulty trials, this interpretation appears to become less likely, as correct responses in high difficulty trick-trials are also errors for the observed person. Our behavioral finding that only the expectancies in low difficulty trials were modulated by empathy instead appears to suggest that expectancy plays a modulatory role for the influence of empathy on the processing of observed actions.

### Empathy, expectation, and the processing of observed actions

Our results show an effect of empathy on both the expectancy data and the ERPs. In an additional exploratory analysis, we examined the relationship between expectation, empathy and the ERP amplitudes in the eight conditions in one LME analysis. We found a positive relationship between expectancy and ERP amplitude, which was not modulated by empathy. We also found that including empathy as a predictor did not explain significantly more variance than using expectancy alone. Keeping in mind that empathy and expectancy were correlated (see above), these results suggest that empathy did not influence ERPs in the present study beyond the effect it had on expectancy formation. This might be an indication that empathy did not directly influence ERPs but via a positive effect on the expectancy formation. Several studies addressed the relationship between trait empathy and the processing of observed actions or action outcomes (Bellebaum et al., 2020; Fukushima & Hiraki, 2006, 2009; Kobza, Thoma, Daum, & Bellebaum, 2011; Newman-Norlund et al., 2009; Shane et al., 2009), but expectancy was rarely taken into consideration, which might be one reason for the inconsistency in the findings. In the context of outcome processing, Lockwood et al. (2015) described that a subregion of the ACC specifically predicted other's rewards in highly empathic participants, whereas activity in that region was comparable for other's and own rewards in low empathizers. These findings may mean that empathy facilitates the generation of predictions based on other's assumed mental states by helping to see an event from the perspective of the observed person. At the same time, positive linear relationships between empathy and expectancy are not always found. In our previous related study (Bellebaum et al., 2020), the observers could more easily predict what action the observed person was about to perform as task difficulty was not varied. Accordingly, participants developed strong expectations regarding the observed person's response with little interindividual variability and thus little room for a modulation by empathy. Brown and Brüne (2012) suggest that predictions in social contexts may depend on similar processes as predictions in nonsocial contexts, but that additional (social) factors play a role only in social contexts. Extending this assumption based on the present findings, it might be that the more the context is dominated by social factors, the more predictions might be modulated by trait empathy. This idea finds some support by findings of Fukushima and Hiraki (2009), who reported that empathy affected the observer FRN only if participants observed humans, not if they observed PC programs.

### Limitations

Due to a relatively large number of exclusions, we analyzed a smaller sample than planned originally. Investigating individual differences in a small sample can lead to false-positive results. LME analysis allows for the inclusion of random effects, so that further interindividual differences besides empathy are at least partly subtracted from the results (i.e., noise is removed; Baayen, Davidson, & Bates, 2008). Nevertheless, it cannot be excluded that the results in the present study represent a false positive result and thus they should be interpreted with caution. Future studies should aim for an increased sample size when investigating effects of empathy on error processing.

## Conclusions

Applying a complex action observation task with true- and false-belief conditions, we found that expectancy, not vicarious errors, was reflected in ERPs time-locked to the observed response, although in an earlier time window as previously suspected. Both the expectancy of the observed action and the ERPs following the observed action were modulated by empathy. We suggest that trait empathy facilitates the processing of stimuli and events from another person's perspective by facilitating expectancy formation. Furthermore, empathy seems to be necessary for expectancy formation only for specific contexts in which social factors dominate. The results found in this study, specifically regarding the indirect influence of empathy on performance monitoring via expectation generation, could help to understand the nature of the problems in social interactions typically found in patients with reduced empathic abilities and may have implications for therapeutic approaches. For example, adding information that makes it easier for these persons to predict and understand others' actions may improve their social skills. Further research needs to investigate the factors that determine the timing of expectancy and empathy modulations in the processing of observed actions.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13415-020-00857-7.

**Acknowledgment** The authors thank Jacqueline Cobilanschi and Clemens Jos Hennig Wallrath for their help with data collection.

**Open Practice Statement** The data and materials for the experiment are available on request (contact christine.albrecht@hhu.de). The experiment was not preregistered.

**Funding** Open Access funding enabled and organized by Projekt DEAL.



## References

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*: Sage.

Alexander, W.H., & Brown, J.W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338-1344.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163-175.

Bates, A.T., Patel, T.P., & Liddle, P.F. (2005). External behavior monitoring mirrors internal behavior monitoring. Error-related negativity for observed errors. *Journal of Psychophysiology*, *19*, 281-288.

Bellebaum, C., Ghio, M., Wollmer, M., Weismüller, B., & Thoma, P. (2020). The role of trait empathy in the processing of observed actions in a false-belief task. *Social cognitive and affective neuroscience*, nsaa009. Advance online publication.

Birch, S., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*, 382-386.

Brown, E.C., & Brüne, M. (2012). The role of prediction in social neuroscience. *Frontiers in human neuroscience*, *6*, 147.

Burke, C.J., Tobler, P.N., Baddeley, M., & Schultz, M. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14431-14436.

Carp, J., Halenar, M.J., Quandt, L.C., Sklar, A., & Compton, R.J. (2009). Perceived similarity and neural mirroring: evidence from vicarious error processing. Social Neuroscience, *4*(1), 85-96.

de Bruijn, E.R., & von Rhein, D.T. (2012). Is your error my concern? An event-related potential study on own and observed error detection in cooperation and competition. *Frontiers in neuroscience*, *3*, 6-8.

de Haen, J. (n.d.). *Deutsche Version der Cambridge Behavior Scale*. Autism Research Centre. http://docs.autismresearchcentre.com/tests/EQ_Deutsch.pdf

Dehaene, S., Posner, M.I., & Tucker, D.M. (1994). Localization of neural system for error detection and compensation. *Psychological science*, *5*(5), 303-305.

Donnarumma, F., Costantini, M., Ambrosini, E., Friston, K., & Pezzulo, G. (2017). Action perception as hypothesis testing. *Cortex*, *89*, 45-60.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and clinical neurophysiology*, *78*(6), 447-455.

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: a tutorial. *Biological Psychology*, *51*(2-3), 87-107.

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.

Flanagan, J.R., & Johansson, R.S. (2003). Action plans used in action observation. *Nature*, *424*(6950), 769-771.

Fukushima, H., & Hiraki, K. (2006). Perceiving an opponent's loss: gender-related differences in the medial-frontal negativity. *Social Cognitive and Affective Neuroscience*, *1*(2), 149-157.

Fukushima, H., & Hiraki, K. (2009). Whose loss is it? Human electrophysiological correlates of non-self-reward processing. *Social Neuroscience*, *4*(3), 261-275.

Gehring, W., Liu, Y., Orr, J., & Carp, J. (2012). The error-related negativity (ERN/Ne). In Luck, S.J. & Kappenman, E.S. (Eds.), *The Oxford Handbook of Event-Related Potential Components* (pp. 231-291). New York, NY: Oxford University Press.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*, 385–390.

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.

Green, P., & MacLeod, C. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution. n/a-n/a. https://doi.org/10.1111/2041-210X.12504.

Holroyd, C.B., & Coles, M.G.H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Reviews*, *109*(4), 679-709.

Koban, L., & Pourtois, G. (2014). Brain systems underlying the affective and social monitoring of actions: an integrative review. *Neuroscience and biobehavioral reviews*, *46*(Pt 1), 71-84.

Koban, L., Pourtois, G., Vocat, R., & Vuilleumier, P. (2010). When your errors make me lose or win: event-related potentials to observed errors of cooperators and competitors. *Social Neuroscience*, *5*(4), 360-374.

Kobza, S., & Bellebaum, C. (2013). Mediofrontal event-related potentials following observed actions reflect an action prediction error. *The European Journal of Neuroscience*, *37*(9), 1435-1440.

Kobza, S., Thoma, P., Daum, I., & Bellebaum, C. (2011). The feedback-related negativity is modulated by feedback probability in observational learning. *Behavioural Brain Research*, *225*(2), 396-404.

Lockwood, P.L., Apps, M.A., Roiser, J.P., & Viding, E. (2015). Encoding of vicarious reward prediction in anterior cingulate cortex and relationship with trait empathy. *The journal of neuroscience: the official journal of the Society for Neuroscience*, *35*(40), 13720-13727.

Loehr, J. D., Kourtis, D., Vesper, C., Sebanz, N., & Knoblich, G. (2013). Monitoring individual and joint action outcomes in duet music performance. *Journal of cognitive neuroscience*, *25*(7), 1049–1061.

Luke S. G. (2017) Evaluating significance in linear mixed-effects models in R. *Behavior research methods*, *49*, 1494–502.

Maier, M.E. & Steinhauser, M. (2016). Error significance but not error expectancy predicts error-related negativities for different error types. *Behavioural Brain Research*, *297*, 259-267.

Miltner, W. H. R., Brauer, J., Hecht, H., Trippe, R., & Coles, M. (2004). Parallel brain activity for self-generated and observed errors. *In Errors, Conflicts, and the Brain. Current Opinions on Performance Monitoring* (pp. 124-129). Leipzig: Max Planck Institute for Human Cognitive and Brain Sciences.

Moreau, Q., Candidi, M., Era, V., Tieri, G., Aglioti, S.M. (2020). Midline frontal and occipito-temporal activity during error monitoring in dyadic motor interactions. *Cortex*, *127*, 131-149.

Newman-Norlund, R.D., Ganesh, S., van Schie, H.T., de Bruijn, E.R., & Bekkering, H. (2009). Self-identification and empathy modulate error-related brain activity during the observation of penalty shots between friend and foe. *Social Cognitive and Affective Neuroscience*, *4*(1), 10-22.

Ninomiya, T., Noritake, A., Ullsperger, M., & Isoda, M. (2018). Performance monitoring in the medial frontal cortex and related neural networks: From monitoring self-actions to understanding others' actions. *Neuroscience research*, *137*, 1–10.

Notebaert, W., Houtman, F., Opstal, F. V., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: an orienting account. *Cognition*, *111*(2), 275–279. https://doi.org/10.1016/j.cognition.2009.02.002

Oliveira, F.T.P., McDonald, J.J., & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations. *Journal of Cognitive Neuroscience*, *19*(12), 1994-2004.

Pezzetta, R., Nicolardi, V., Tidoni, E., & Aglioti, S.M. (2018). Error, rather than its probability, elicits specific electrocortical signatures: a combined EEG-immersive virtual reality study of action observation. *Journal of Neurophysiology*, *120*(3), 1107-1118.

Ridderinkhof, K.R., Ullsperger, M., Crone, E.A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*(5695), 443-447.

Schiffer, A.M., Krause, K.H., & Schubotz, R.I. (2014). Surprisingly correct: unexpectedness of observed actions activates the medial prefrontal cortex. *Human Brain Mapping*, *35*(4), 1615-1629.

Shane, M.S., Stevens, M.C., Harenski, C.L., & Kiehl, K.A. (2009). Double dissociation between perspective-taking and empathic-concern as predictors of hemodynamic response to another's mistakes. *Social Cognitive and Affective Neuroscience*, *4*(2), 111-118.

Steinhauser, M., & Andersen, S. K. (2019). Rapid adaptive adjustments of selective attention following errors revealed by the time course of steady-state visual evoked potentials. *NeuroImage*, *186*, 83–92.

Taylor, S.F., Stern, E.R., & Gehring, W.J. (2007). Neural systems for error monitoring: recent findings and theoretical perspectives. *The Neuroscientist*, *13*(2), 160-172.

Ullsperger, M., Danielmeier, C., & Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiological reviews*, *94*(1), 35–79.

van Schie, H.T., Mars, R.B., Coles, M.G., & Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, *7*(5), 549-554.

Wang, L., Tang, D., Zhao, Y., Hitchman, G., Wu, S., Tan, J., & Chen, A. (2015). Disentangling the impacts of outcome valence and outcome frequency on the post-error slowing. *Scientific Reports*, *5*, 8708.

Wellman, H.M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false believe. *Child Development*, *72*(3), 655-684.

Wessel J. R. (2018). An adaptive orienting theory of error processing. *Psychophysiology*, 55(3), https://doi.org/10.1111/psyp.13041.

Wessel, J.R., Danielmeier, C., Morton, J.B., & Ullsperger, M. (2012). Surprise and error: common neuronal architecture for the processing of errors and novelty. *The journal of neuroscience*, *32*(22), 7528-7537.

Yeung, N., Holroyd, C.B., & Cohen, J.D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex*, *15*(5), 535-544.

Received: 9 December 2020 | Revised: 5 May 2021 | Accepted: 1 June 2021

DOI: 10.1111/psyp.13883

ORIGINAL ARTICLE

PSYCHOPHYSIOLOGY SPR WILEY

# Disentangling effects of expectancy, accuracy, and empathy on the processing of observed actions

**Christine Albrecht** | **Christian Bellebaum**

Institute of Experimental Psychology, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

**Correspondence**
Christine Albrecht, Institute of Experimental Psychology, Heinrich Heine University Düsseldorf, Universitätsstraße 1, Building 23.03, Room Number 00.89, 40225 Düsseldorf, Germany.
Email: christine.albrecht@hhu.de

**Abstract**

A number of studies suggest that event-related potential (ERP) components previously associated with error processing might represent expectation violation instead of valence. When observing others, these processes might further be modulated by trait empathy. We suggest that trait empathy modulates expectancy formation and that these expectancies then influence observed response processing as reflected in a frontocentral negative ERP component resembling the previously described observer error-related negativity. We acquired single trial ERPs of participants who observed another person in a true- or false-belief condition answering correctly or erroneously. Additionally, we prompted participants' expectancy in some trials. Using linear mixed model analyses, we found that for low empathy participants, expectations for the false-belief condition decreased throughout the experiment, so that expectations were more pronounced in participants with higher empathy toward the end of the experiment. We also found that single trial expectancy measures derived from regression models of the measured expectancies predicted the amplitude of the frontocentral negative ERP component, and that neither the addition of empathy nor accuracy or trial type (true- or false-belief) led to the explanation of significantly more variance compared with the model just containing expectancy as predictor. These results suggest that empathy modulates the processing of observed responses indirectly via its effect on expectancy of the response.

**KEYWORDS**

action observation, EEG, empathy, error processing, expectation, oERN

## 1 | INTRODUCTION

Error processing in the human brain has been studied extensively in the past 30 years. Functional magnetic resonance imaging (fMRI) studies as well as studies using electroencephalography (EEG) and source analysis have proposed a close link between error commission and activity in the posterior medial frontal cortex (pMFC; Ullsperger et al., 2014) and anterior cingulate cortex (ACC; Dehaene et al., 1994; Holroyd et al., 2004; Ullsperger & Cramon, 2004; for reviews, see Ridderinkhof et al., 2004; Taylor et al., 2007). Frontocentral activity seems to be reflected in a specific event-related potential (ERP) component (Dehaene et al., 1994; Ridderinkhof et al., 2004), characterized by a negative peak around

100 ms after error response onset at frontocentral electrodes, called the error negativity or error-related negativity (ERN; Falkenstein et al., 1991, 2000; Gehring et al., 1993; Hajcak et al., 2005; for reviews, see Gehring et al., 2012; Holroyd & Coles, 2002). More recent studies have suggested that this effect is not limited to self-committed errors (for a review, see Koban & Pourtois, 2014). Research investigating neural correlates of the processing of observed errors in others also found an involvement of the pMFC and ACC (de Bruijn et al., 2009; Desmet et al., 2014; Ninomiya et al., 2018; Shane et al., 2008), although also other brain regions such as the superior temporal sulcus were activated in addition (Ninomiya et al., 2018), as well as a similar negative ERP component (Miltner et al., 2004; van Schie et al., 2004; for a review, see Koban & Pourtois, 2014). While the topography and origin of this so-called observer error-related negativity (oERN) were comparable with the ERN (Miltner et al., 2004; van Schie et al., 2004), the oERN has a smaller amplitude (Miltner et al., 2004; van Schie et al., 2004) and a longer peak latency, with a larger variance between studies (130–300 ms; Bates et al., 2005; Koban et al., 2010; as opposed to Carp et al., 2009; de Bruijn & von Rhein, 2012; Miltner et al., 2004), possibly depending on task complexity (Koban & Pourtois, 2014).

In most experimental settings investigating (observed) error processing, errors are less likely than correct responses. This is in accordance with everyday experience where errors, at least for easy or routine tasks, are generally infrequent. However, these study designs make it impossible to disentangle the processing of expectation violations from error processing. Recent evidence has emphasized the role of predictions and prediction errors for ACC activity, indicating that the ACC, and consequently also those ERP components with an ACC origin that were previously associated with error processing, may code for unexpected events rather than errors (Alexander & Brown, 2011; Desmet et al., 2014; Ferdinand et al., 2012; Jessup et al., 2010; Oliveira et al., 2007; Schiffer et al., 2014; Wessel et al., 2012, 2014; Zubarev & Parkkonen, 2018). While watching the actions of others, observers seem to form predictions (Donnarumma et al., 2017; Flanagan & Johansson, 2003; Friston et al., 2012). If these predictions are not met, this results in an action prediction error (Brown & Brüne, 2012; Burke et al., 2010), which has been suggested to drive frontocentral activity (Desmet et al., 2014; Schiffer et al., 2014) and elicit an oERN-like component (Albrecht & Bellebaum, 2021; Bellebaum et al., 2020; Kobza & Bellebaum, 2013; Wang et al., 2015). In the following, we will refer to this component as frontocentral negative ERP component as its exact function is still to be determined.

Action observation takes place in a social context. It is thus conceivable that social cognitive skills such as empathy play a role in the processing of another's actions. A range of studies provided support for effects of empathy processes in action observation by showing a link between state empathy and error monitoring (Carp et al., 2009; de Bruijn & von Rhein, 2012; Kang et al., 2010; Koban et al., 2012; Marco-Pallarés et al., 2010; Mobbs et al., 2009; Weller et al., 2018). In contrast to that, studies investigating the role of trait empathy had inconclusive results (Brazil et al., 2011; Clawson et al., 2014; Lockwood et al., 2015; Newman-Norlund et al., 2009; Shane et al., 2009).

This could partially be explained by assuming that the frontocentral negative ERP component does not reflect observed error processing, but the processing of prediction errors (see Alexander & Brown, 2011) and thus unexpectedness or surprise of the observed action (Bellebaum et al., 2020; Kobza & Bellebaum, 2013; Wang et al., 2015). In this respect, empathic processes may facilitate the formation of expectancies concerning the observed action (Brown & Brüne, 2012), which is a prerequisite for prediction error processing. In how far the generation of predictions depends on empathic processes, however, might then be task-dependent: less or no empathy might be required when others' actions are relatively straight-forward (e.g., Clawson et al., 2014). For complex social tasks like false-belief-tasks, however, empathy may be needed to form predictions (Albrecht & Bellebaum, 2021; Bellebaum et al., 2020; Birch & Bloom, 2007; Ferguson et al., 2015; Wellman et al., 2001).

In two previous studies investigating the effect of empathy on action observation (Albrecht & Bellebaum, 2021; Bellebaum et al., 2020), we applied a false-belief-task paradigm in the form of a two-shell game (Kobza & Bellebaum, 2013), as we hypothesized that empathy is particularly important in tasks with a false-belief (Birch & Bloom, 2007; Wellman et al., 2001). In this paradigm, participants observe a player who has to decide which of two shells contains a hidden ball. In some trials, the player is tricked and the ball is swapped between the shells, but only the observer participant has access to this information, so that this condition entails a false belief of the player. This experimental manipulation aims to induce two expectancy conditions: observer participants should expect the player to answer correctly in no-trick trials, but erroneously in trick trials. Blocks of experimental trials were interspersed with prompt trials in which expectancies of the observers were assessed. In Bellebaum et al. (2020), we found that a frontocentral negative ERP component in the typical oERN time window reflected the induced expectancies, and that higher trait empathy enhanced this effect. The results were discussed in terms of a facilitation of expectancy formation via empathic abilities, but there was no significant relationship between empathy and the strength of the expectancy concerning the observed response, so that the mechanism by which empathy affects the processing of observed responses remained unclear. An alternative explanation could be that observers processed errors from the observed person's point of view, with enhanced amplitudes in conditions in which the observed response was a subjective error for the observed person, who didn't know that he/she was tricked.

In a follow-up study (Albrecht & Bellebaum, 2021), we thus tried to disentangle the two alternative explanations by adding the factor task difficulty to the paradigm, because an (observed) error in a more difficult task may be more expected than an error in an easy task, but it remains an error from both the player's and observer's perspective. We indeed found first evidence that empathy affects expectancy, which drives the processing of observed responses. Possibly due to the higher complexity of the task and subsequent changes in the procedure (block design) an earlier frontocentral negativity compared with Bellebaum et al. (2020) was modulated, around the N1 component. Although the results were thus not directly comparable, the study by Albrecht and Bellebaum (2021) revealed that adding variability concerning the expectancy of the observed response can be a promising approach in studying the relationship between empathy, expectancy, and the processing of observed responses. In the present study, we thus aimed to modify our original study to increase the variability of expectancies and examine their effect further.

In Bellebaum et al. (2020), we analyzed a sample of 50 women, who were mostly psychology students. The homogeneity of the sample might have contributed to the ceiling/floor effect in expectancies with low variance (for a display, see supplementary material, Figure S2), which might be one reason that the study found no effect of empathy on expectancy formation. We reasoned that studying the relationship between individual continuous predictor variables (expectancy, empathy) and observed response processing calls for a rather large and more heterogeneous sample with variability in the variables of interest. We thus decided to extend the already acquired sample by recruiting a comparably sized sample of men from various backgrounds. This had the desired effect of increasing the variability of expectancies (see Figure S2).

The second change with respect to the study by Bellebaum et al. (2020) is of methodological nature. Expectancies may have changed during the experiment, and therefore trial-by-trial variations of expectancy were taken into account, as has, for example, also been done in studies on reward prediction error processing (Burnside et al., 2019). One factor contributing to the change in expectancies could be that correct and error responses by the player were counterbalanced. During the course of the experiment, participants may thus have adapted their expectancies based on the actual probabilities of the responses, and this adaptation may also differ between participants. To elucidate the relationship between empathy, expectancy, and observed response processing, we thus computed single-trial values for observer participant's expectancies and included the resulting expectancy values into an ERP analysis based on single trial data. We hypothesized that expectancy formation would be dependent on trait empathy, possibly modulated by changes over the experimental course. Furthermore, and in accordance with a previous related study (Albrecht & Bellebaum, 2021), we expected that trial-by-trial variations in expectancy would affect single-trial ERP data of a frontocentral negative component previously associated with action observation, but that trait empathy would not explain any further variance in this model.

# 2 | METHOD

## 2.1 | Participants

We acquired data from a sample of 105 participants. In a recent study on the relationship between empathy and the monitoring of observed responses we have already reported an analysis based on 50 of these participants, all women (Bellebaum et al., 2020). As outlined in the Introduction, our main interest in the present study was to find out about the interplay between empathy, expectancy, and observed action processing. As the original, quite homogeneous sample (only women, mostly from the same background as psychology undergraduates) showed low variability in the expectancy measures (a ceiling/floor effect emerged in the no-trick and trick conditions, respectively), we decided to compile a more heterogeneous sample to add variability, which should also affect expectancy measures. For this purpose, we added 55 men from various backgrounds to the sample, so that the overall sample was balanced with respect to gender. Indeed, the variability of expectancy values was increased (see supplementary material, Figure S2), which meant a reduced ceiling/floor effect. Note that the men were added to enhance the original sample variability, not as a second group, and thus the men and women groups were not systematically matched concerning age (see Table S1). While this was advantageous regarding sample heterogeneity, we refrained from a direct comparison between men and women in this study. Three participants (one woman and two men) had to be excluded because for one the ERP component of interest could not be identified and two failed to answer to the prompt questions with which we assessed the expectancy of the observed response (see below). Of the remaining 102 participants, 53 were men and 49 women. All participants were between 19 and 38 years old ($M = 25.6$, $SD = 4.4$). None of the participants reported a history of neurological diseases or the acute intake of drugs or medication that would affect the nervous system. All participants had normal or corrected-to-normal vision. Participants received 15 € or course credit as reimbursement. The study was approved by the ethics committee of the Faculty of Mathematics and Natural sciences at Heinrich-Heine-University, Düsseldorf, and is in accordance with the declaration of Helsinki.

## 2.2 | Experimental task

We applied a Two-Shell Game paradigm identical to that described in Bellebaum et al. (2020) and Kobza and Bellebaum

(2013). In this paradigm, the participants observe the actions of a fictitious player who ostensibly watches a ball being hidden under one of two shells and, after a few rotations, has to decide under which shell he/she suspects the ball. The player was introduced with a name and a picture on the screen at the beginning of the experiment. During the experiment, observer participants saw the shells, ball, and hand of the player (holding a joystick) from above. The trials were simulated, with the player answering correctly and erroneously equally often, but observer participants were told that they observed a recording of an actual player. To facilitate empathising, all women observed a (fictitious) female player while all men observed a (fictitious) male player.

Observer participants were told that the player observed the game from the front and would use a joystick pointing either to the left or the right to select the specific shell. The observer participants, however, saw the game from above and were thus also able to see the ball throughout the whole game. To modify the observer participant's expectations toward the player's answer, a false-belief condition was added. Observer participants were told that the player was tricked in half of the trials, where the ball was swapped between the shells. This would be almost never visible to the player, while observer participants saw the trick. The starting point of the ball (right or left), the number of rotations (two to four) and, respectively, the end point of the ball (right or left) were balanced throughout the experiment. In trick trials, it was also balanced during which rotation the ball was swapped. In total, participants observed 468 trials, divided equally into four blocks between which observer participants could take short breaks. There were four experimental conditions: no-trick trials followed by either the correct or erroneous answer and trick-trials followed by either the correct or erroneous answer. According to the expectations set in the instructions, observer participants should expect correct answers in no-trick and errors in trick trials. All conditions occurred equally often and in random order throughout the experiment. Of all trials, 48 (24 of each trial type, trick and no-trick) ended not in the player answering, but in the prompt question: "Where will the player point the joystick?" which the observer participants were instructed to answer with a respective left or right key. These prompt questions were spread throughout the experiment. This ensured that the observer participants were paying attention and provided a measure of each participant's actual expectancies. For the time course of events in each trial, please refer to Figure 1.

## 2.3 | Empathy measure

As an empathy measure, we selected the Empathy Quotient (EQ) obtained with the German version of the Cambridge Behavior Scale (Baron-Cohen & Wheelwright, 2004; de Haen, n.d.). This questionnaire consists of 60 items, of which 20 are distractors. The items are statements that participants can agree or not agree to on a 4-point Likert scale (from "strongly disagree" to "strongly agree"). Items are negatively or positively scaled. A maximum of 2 points can be obtained per item: for positively scaled items, 2 for "strongly agree," 1 for "slightly agree" or 0 for "slightly disagree" or "strongly disagree." For negatively scaled items, the scoring is reversed. No points can be obtained for the distractor items. All points are added up and result in an EQ sum sore between 0 and 80. In the present study, empathy scores ranged from 23 to 71 ($M = 44.4$, $SD = 9.7$) and data were normally distributed (see Figure S1).

The EQ was chosen as this general measure of empathy, including both affective and cognitive aspects of empathy, affected processing in the particular task used in this study (Albrecht & Bellebaum, 2021; Bellebaum et al., 2020) and in another false-belief task (Ferguson et al., 2015). In the female sample, the German version (Paulus, 2009) of the Interpersonal Reactivity Index (IRI; Davis, 1980, 1983) was considered as well (Bellebaum et al., 2020), but no significant modulation of the ERPs time-locked to the observed response was found for the IRI or its subscales of cognitive or affective empathy, possibly due to the lower number of items and the reduced variability in the obtained scores (Koller & Lamm, 2014).

## 2.4 | Procedure

After arrival in the lab, all participants gave informed written consent to take part in the study. They then completed a demographic questionnaire as well as two empathy questionnaires. In addition to the German version of the Cambridge Behavior Scale, participants completed a German short version (Paulus, 2009) of the IRI (Davis, 1980, 1983). As mentioned above, we did not consider the IRI in the present study.

After the questionnaires were completed, the EEG electrodes were applied to the scalp and participants were sat in front of a desktop monitor with a screen resolution of 1,920 × 1,080 px, on which the computer experiment was presented. This took about 45 min. Stimulus presentation and the recording of the responses were controlled by means of the software Presentation (Version 20.0, Neurobehavioral Systems Inc).

## 2.5 | EEG recording

EEG signals were recorded with active silver/silver-chloride electrodes with a sampling rate of 1,000 Hz. The electrodes were attached according to the 10–10 system on 29 scalp sites which were F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6,

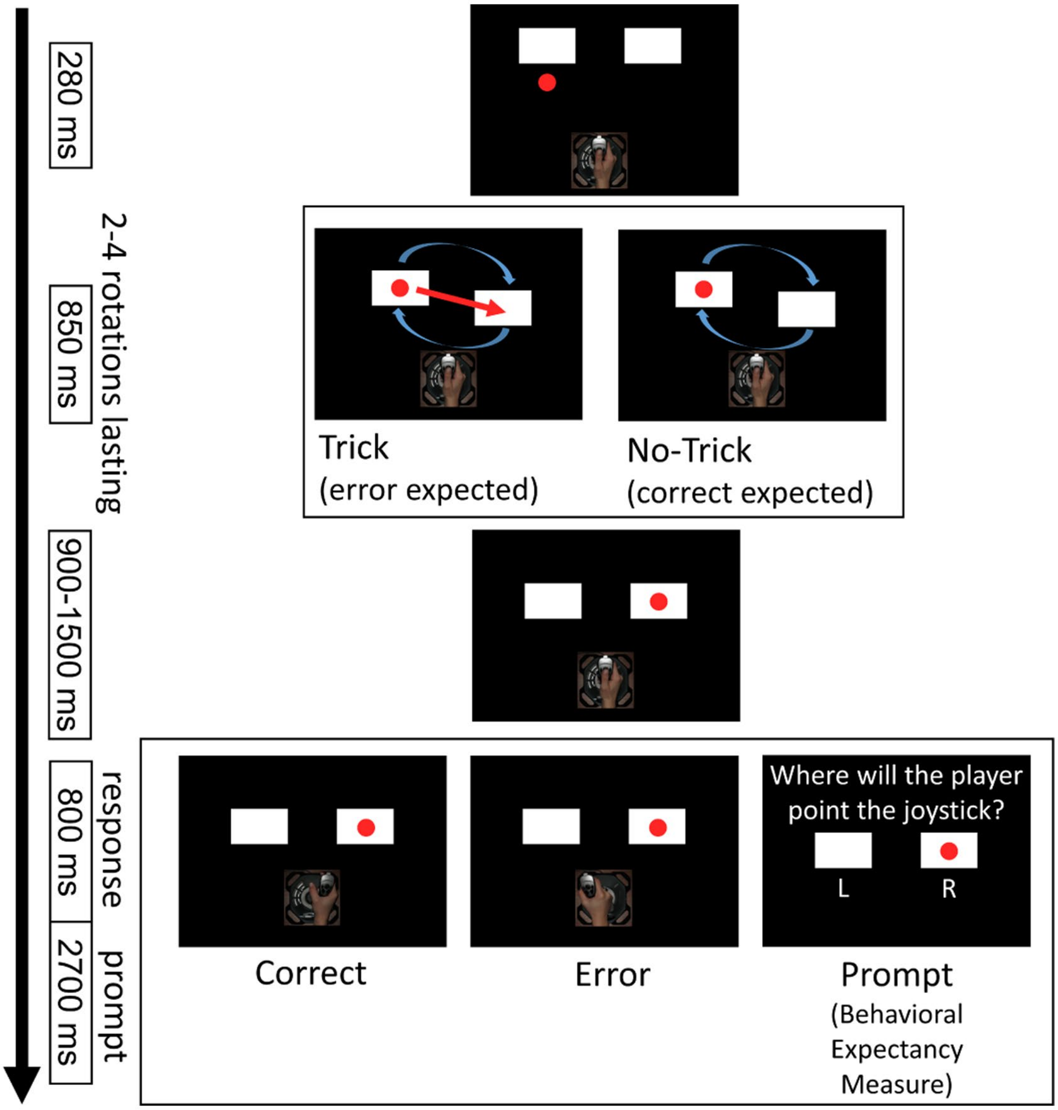



**FIGURE 1** Time course of events in the experiment trials. The observer participants watched trick or no-trick trials that ended in either correct or erroneous answers of the player. The shells changed places several times. In trick trials, the ball was swapped between shells. All shells and ball movements were displayed as a video for the observer participants. In this figure, the movements are marked by arrows. All trials (trick and no-trick) ended either in a response by the observed player, or, in 48 trials, in a prompt question to measure the observer's expectancies. To ensure participants' attention and enhance plausibility, observed response times slightly varied. For details on the experimental conditions please refer to the main text

T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO9, P1, Pz, P2, PO10, and FCz, which was used as online reference. We also recorded EEG signals on both mastoids for re-referencing in the course of data analysis (see below). The ground electrode was placed at AFz. Moreover, we recorded an electrooculogram (EOG) to detect horizontal (hEOG) and vertical (vEOG) eye movements. Two hEOG electrodes were placed at F9 and F10, respectively. One vEOG electrode was placed at Fp2, the other one below the right eye. We used a 32-channel actiCap electrode cap (ActiCAP; brain Products GnbH, Germany) and the signal was recorded with BrainVision Recorder software, version 1.20 (Brain Products, Munich, Germany). Impedances were kept below 10 kΩ.

## 2.6 | Data analyses

### 2.6.1 | Behavioral data

Behavioral data were read from the Presentation logfiles and brought into the correct format to be analyzed in R by using MATLAB, version R2017b (Mathworks, Natick, Massachusetts, USA). We then conducted a single-trial linear mixed models (LME) analysis in R (version 3.5.3). The dependent variable was set as the expected accuracy in each of the 48 trials that served the assessment of expectancies. Although the dependent variable was binary (expected correct or error response), we chose to code it not as 0 and 1, but as 0 and 100, to yield larger *b*-values in the LME analyses (see below) which are easier to interpret, and to derive trial-by-trial expectancy values in percent that could be used for the single-trial ERP analysis (see also below). As independent variables, we included the categorical within-subject factor Trial Type (coded as −1 = no-trick, 1 = trick for the LME) and the continuous between-subject factor Empathy which consisted of the mean-centered EQ scores of the participants. Furthermore, we assumed that expectancies might change during the experiment—mainly because the actual number of correct and erroneous answers did not match the expectancies set during the instructions (mostly correct in no-trick and mostly erroneous in trick trials). Thus, we added Trial Number as another independent variable. As there were 468 trials in the experiment, and the prompt questions were randomly distributed across the trials, Trial Number was a continuous variable reflecting the course of the experiment. For each participant there was a maximum of 48 data points (all answered prompt questions; 24 per Trial Type). Trials in which participants did not answer fast enough were not included in the data analysis. On average, 46.6 data points were included for each participant (*SD* = 2.0). The Trial Number score ranged from 1 to 468. The factor Trial Number was mean-centered for the analysis. Random intercepts were included for participants as well as random slopes of Trial Type and Trial Number per participant.

### 2.6.2 | EEG data

*EEG preprocessing*

We prepared to conduct two analyses of the amplitude of the frontocentral negative component: First, we attempted to replicate the results reported in Bellebaum et al. (2020) in the larger sample and with single-trial EEG data, including the factors Accuracy (correct or incorrect answer), Trial Type (trick or no-trick trial), and Empathy (EQ score). Second, we analyzed the effect of expectancy. For this model, we replaced the factors Accuracy and Trial Type of the first analysis that aimed to affect the expectancy of the observed response, with the factor Expectancy, using the single trial values yielded by a regression model (see below). The EEG data were preprocessed with BrainVision Analyzer (Brain Products, Munich, Germany) and MATLAB (Mathworks, Natick, Massachusetts, USA) in the same way as described by Bellebaum et al. (2020). We applied a 20 Hz low-pass and a 0.5 Hz high-pass filter on the raw data. Subsequently, we performed an independent component analysis on the EEG data of each participant. We selected one component that represented vertical eye movements and blinks, as suggested by its topography (a symmetrical frontal distribution). This component was removed by ICA back-transformation. The data were then segmented into 800 ms epochs that started 200 ms before the player's response (the joystick pointing to the left or to the right), and the pre-response period was used for baseline correction. This was in accordance with Bellebaum et al. (2020). We then ran an automatic artifact rejection that removed all segments with a voltage step larger than 50 μV per ms, all segments where the highest and lowest data point were more than 100 μV apart and all segments in which the signal was higher than 100 μV or lower than −100 μV. The artifact rejection removed on average 5% (*SD* = 6%) of segments. No participant lost more than 30% of trials. The data from all included segments of each participant were exported as text file, as the analyses were based on single trial ERPs. However, we also calculated the average EEG signal for each participant in each of the four conditions trick correct, trick error, no-trick correct, and no-trick error. This was done because the extraction of values for the ERP component of interest for each trial was based on individual participant's average ERPs (see below).

All further analyses were performed on the pooled signal across electrodes in a fronto-central electrode cluster consisting of Fz, FC1, FCz, FC2, and Cz in accordance with Bellebaum et al. (2020). At first, the latencies of the maximum negative peak between 250 and 420 ms and of the preceding positive peak between 130 ms and the negative peak in individual participants' averages were determined (see Bellebaum et al., 2020). We decided to calculate latencies for each condition (trick correct, trick error, no-trick correct, no-trick error) and each participant, as latencies varied not only between conditions for each participant (average standard deviation of the negative peak per participant = 26.8, for the preceding positive peak = 19.4 ms) but also within conditions across participants (negative peak: trick correct: *M* = 325.2 ms, *SD* = 41.1 ms; trick incorrect: *M* = 322.6 ms, *SD* = 45.2 ms, no-trick correct: *M* = 326.1 ms, *SD* = 42.3 ms, no-trick error: *M* = 328.1 ms, *SD* = 42.6 ms; preceding positive peak: trick correct: *M* = 259.2 ms, *SD* = 33.1 ms; trick incorrect: *M* = 254.1 ms, *SD* = 33.8 ms, no-trick correct: *M* = 254.8 ms, *SD* = 32.9 ms, no-trick error: *M* = 258.6 ms, *SD* = 27.1 ms). To obtain amplitude data for single trials, the amplitudes at the time points of individual participant's peak latencies were obtained for each individual segment. Then, peak-to-peak values were calculated by subtracting the maximum positive peak amplitude from the negative peak amplitude. This allowed comparability with analyses of previous studies based on average ERP amplitudes that used a peak-to-peak amplitude approach (Bellebaum et al., 2020).

Before the statistical analysis was performed, outliers were excluded both within and between subjects: We excluded trials in which the peak-to-peak amplitude differed by more than two standard deviations from the respective participant's mean in this condition. On average, 5% of trials were excluded for each participant (*SD* = 4%). Furthermore, two participants for whom the peak-to-peak mean amplitude differed by more than two standard deviations from the group mean in more than one condition were excluded. The remaining participants entering the EEG analyses (48 women, 52 men) were between 19 and 38 years old ($M = 25.5$, $SD = 4.4$). On average, 379.7 trials ($SD = 32.8$; of a maximum of 420 trials) per participant were included in the analyses.

*EEG analysis Accuracy × Trial Type × Empathy*

As a first step, we decided to calculate an LME model that was equivalent to that used in Bellebaum et al. (2020) with the only difference that we used single-trial data. This analysis was conducted because we were interested to see if we could replicate our previous finding of a dissociation in observed response processing between a false- (trick) and a true-belief condition (no-trick) in a sample with more variable and, on average, less pronounced expectations concerning the observed response. The peak-to-peak amplitude in each segment was defined as dependent variable. As independent variable, we defined the categorical factor Accuracy (1 = correct, −1 = error) as well as the categorical factor Trial Type (1 = trick, −1 = no-trick), and the continuous mean-centered factor Empathy. Accuracy and Trial Type were included as random effects for each participant in the model, in addition to random intercepts.

*EEG analysis expectancy*

In a second model we aimed to examine the relationship between expectancy and response-locked ERPs directly. For this purpose, we applied a regression model to derive single-trial values for expectancy.

*Regression model.* As the behavioral data analysis did in fact reveal individual differences in expectancies depending on Trial Number and Trial Type (see Results section for details), we aimed to add trial-by-trial expectancy values (rather than the categorical factors Trial Type and Accuracy) as independent variable in the subsequent EEG analysis in order to examine the impact of expectancy on the frontocentral negative ERP component (see below), taking into account also changes in the expectations across the experiment. For some participants, for whom expectancies stayed the same throughout the experiment, we would expect a course parallel to the x-axis. For other participants, expectancies might be strong at the beginning and become less strong throughout the experiment or the other way around. Expectancies were measured with binary questions (observer participants stated which shell they believed the player would choose) to make answering relatively easy and keep prompt trials similar to observation trials. However, we assume that expectancies and predictions, as has been shown in studies investigating prediction errors (Burnside et al., 2019; Sutton & Barto, 1998), lie on a continuum between not expecting an event at all (coded as 0) or being absolutely sure the event will occur (coded as 1 or 100%). In order to account for the continuity of the underlying variable, we decided to treat answers (0 or 100) as two ends of a scale rather than as a binary variable. As datapoints per trial type were relatively few, we opted for the simplest possible model to display expectancies, namely a linear regression model. In the first step, we thus calculated two linear regression models for each participant, one for trick and one for no-trick trials. The models were calculated with the MATLAB function fitlm (Mathworks, Natick, Massachusetts, USA) and were based on the participants' answers to the prompt questions, representing the linear development of expectancies throughout the experiment. We obtained a formula for each participant and each Trial Type of the type:

$$E = b_0 + b_1 \times t.$$

In this formula, $E$ reflects the expectancy at each time point (in each trial; in %), $b_0$ and $b_1$ represent intercept and slope of the linear model, and $t$ is the time, or more specifically, the trial number. As the question that the participants answered in prompt trials assessed the expectancy of a correct response, the formula allowed us to calculate this expectancy for each of the 420 observation trials. For observed error responses, we calculated the inverted value ($100 - E$), however, so that expectancy measures that entered this analysis always represented the expectancy of the observed response. These expectancy measures were then used as an independent variable to predict single-trial ERP amplitudes of the frontocentral negative component, time-locked to the observed response (see below). This was done to examine the hypothesis that the ERP component was driven by trial-to-trial variations of expectancy.

*LME model.* The LME model used for this calculation included Expectancy as a mean-centered continuous factor. By this, we replaced the factors Accuracy and Trial Type, that in our previous study (Bellebaum et al., 2020) served as factors that affect expectancy, with the actual, measured expectancy values of each participant. As expectancy values already varied between participants, Expectancy was not included as a random effect in this model, but we included random intercepts and slopes per participant.

*Effect of Accuracy, Trial Type, and Empathy on model's fit*

In the previous analysis, only expectancies were used to predict the amplitude of the frontocentral negative ERP component. To determine if one of the factors used in the first LME EEG analysis (Accuracy, Trial Type, and Empathy) explained additional variance above the variance explained by Expectancy (and thus to investigate the hypotheses that trait empathy would not explain any further variance in this model), we compared models with chi square tests: The Expectancy

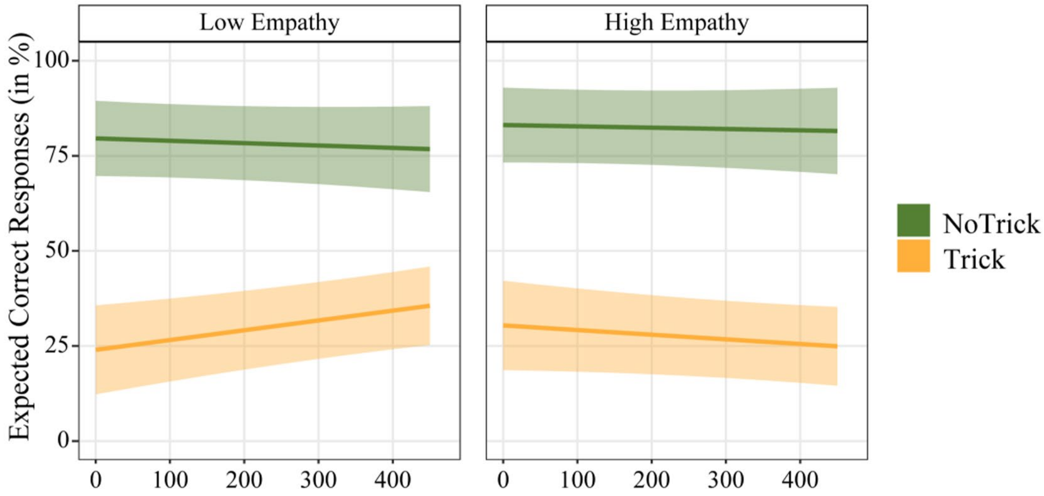



**FIGURE 2** Behavioral expectancy data as a function of Trial Type, Empathy and Trial Number. Confidence intervals are displayed around the regression lines

model with, respectively, an Expectancy × Accuracy, an Expectancy × Trial Type, and an Expectancy × Empathy model. Coding and random effects remained the same as in the previous analyses; for the Expectancy × Accuracy model and the Expectancy × Trial Type model, the categorical factor was included as random factor. To make a comparison possible, the models were calculated with a maximum likelihood instead of a restricted maximum likelihood approach as used for the Satterthwaite's approximation.

### 2.6.3 | Interactions

For resolving significant interactions, we calculated conditional slopes, meaning the slope of a specific effect when one predictor was held constant. For categorical factors, we used either −1 or 1 as constants, according to the coding of the respective variable. For continuous factors, we investigated effects of the remaining factors at the mean value of the continuous factor minus one standard deviation ($M - 1$ *SD)* and, correspondingly, at $M + 1$ *SD* which allowed us to investigate effects for low and high levels of the respective continuous factor. We resolved significant interactions in a step-wise manner: while holding one factor constant, we checked for significant lower-level interaction effects and iterated this procedure until all factors were resolved.

# 3 | RESULTS

In the following, the main results for the behavioral and EEG data are presented. Please find additional statistical data for each LME analysis in the supplementary material (Tables S2–S4).

## 3.1 | Behavioral data

On average across the whole experiment, observer participants expected the player to answer correctly in 80.0% (*SD* = 19.8%) of the prompt trials in no-trick trials. In trick trials, observer participants expected the player to answer correctly in 28.9% (*SD* = 25.1%). The expectancies were thus less strong and more variable than in the study by Bellebaum et al. (2020), as was intended in the present study. Nevertheless, the expectancy difference between conditions was highly significant, $t(101) = 13.17$, $p < .001$, suggesting that the instruction was successful in inducing different expectations concerning the accuracy of the observed action in the trick and no-trick conditions. As described in the Methods section, a behavioral data analysis by means of LME was conducted to examine the development of expectancy in the trick and no-trick conditions across the experiment, and in how far the expectancy was affected by empathy.

Figure 2 shows the relationship between the behavioral expectancy measure, Trial Number, and Empathy according to Trial Type (trick and no-trick condition). In accordance with the result reported above, the analysis revealed a main effect of Trial Type, $F(1, 174.40) = 56.02$, $p < .001$, $b = -25.71$, indicating smaller expected accuracy for trick than for no-trick trials. No other significant main effects were found (all $p \geq .726$). The analysis further revealed a significant three-way interaction between Trial Type, Trial Number, and Empathy, $F(1, 4{,}807.40) = 6.38$, $p = .012$, $\eta_p^2 < .01$. To further resolve this interaction, we examined effects of the other factors for low and high levels of Empathy. For participants with lower empathy, we found a significant two-way interaction of Trial Type and Trial Number, $F(1, 4{,}814.20) = 7.85$, $p = .005$, but no such effect for participants with higher empathy ($p = .441$). A significant main effect of Trial Number for low Empathy participants emerged in trick trials, $F(1, 195.40) = 7.99$, $p = .005$, $b = .03$, but not in no-trick trials ($p = .490$). This indicates that the expectancy of correct responses increased in trick trials for participants with low empathy. Figure S3 provides an alternative way to illustrate the three-way interaction. Here it can be seen that, descriptively, higher empathy participants show a larger difference in expectancies between trick and no-trick trials toward the end of the experiment, which resembles the pattern of expectancies we found in Albrecht and Bellebaum (2021), where we did not examine the development of expectancies over time (see also Figure S4). There were no other significant interaction effects for the analysis of expectancies across the experiment (all $p \geq .085$).

## 3.2 | EEG analysis: Accuracy × Trial Type × Empathy

Grand average ERPs for the four conditions trick correct, trick error, no-trick correct, no-trick error, and the respective topographies as well as the topographies of the error-correct difference waves are displayed in Figure 3. Due to the high interindividual peak latency variability the between-condition differences might be underestimated in the grand-average ERPs. Nevertheless, it can be seen that the peak-to-peak amplitude of the frontocentral negativity differs between observed correct and error responses for no-trick, but not for trick trials, which becomes even more apparent in the small display of the amplitudes and in the bar plots showing peak-to-peak amplitudes. Descriptive data for single-trial peak-to-peak amplitudes of the frontocentral negative component according to Trial Type, Accuracy, and Empathy are displayed in Figure 4. In the Accuracy × Trial Type × Empathy LME model, we found a three-way interaction between all factors, $F(1, 37{,}789.00) = 12.59$, $p < .001$, $\eta_p^2 < .01$. When resolving for empathy, a significant interaction effect of Accuracy and Trial Type emerged only for high empathy participants, $F(1, 37{,}787.00) = 14.60$, $p < .001$, but not for low empathy participants ($p = .231$). The slope for Accuracy was significant in no-trick trials, $F(1, 611.00) = 8.24$, $p = .004$, $b = .27$, as well as trick trials, $F(1, 620.00) = 5.88$, $p = .016$, $b = -.23$. In no-trick trials, amplitudes were more negative after observing errors than correct responses; the opposite was true for trick trials in high empathy participants. This is consistent with the descriptive pattern of expectancies, which showed more pronounced differences between trick and no-trick trials for high than low empathy participants (for an additional display, see Figure S4), although this was most pronounced toward the end of the experiment (see Figure S3). The LME analysis revealed a trend for an overall main effect of Accuracy, $F(1, 162.00) = 3.54$, $p = .062$, $b = .09$, $\eta_p^2 < .01$ (amplitudes were more negative for error responses), but no other main effects (all $p \geq .438$). We also found a trend for the Trial Type × Accuracy interaction, $F(1, 37{,}786.00) = 3.44$, $p = .064$, $\eta_p^2 < .01$. As this was only a trend and we also found a higher-order interaction, we did not resolve this interaction. No other interactions were found (all $p \geq .148$).

## 3.3 | EEG analysis: The effect of expectancy

Please refer to Figure 5 for the descriptive data of the analysis of the frontocentral negative ERP component with Expectancy as independent variable. For a scatterplot containing individual data points, please refer to the supplementary material (Figure S5). We found a significant effect of Expectancy, $F(1, 37{,}876.00) = 12.14$, $p < .001$, $b < .01$, $\eta_p^2 < .001$. The more unexpected the observed response was, the more negative was the amplitude of the negative frontocentral component.

We then performed three separate Chi-Square tests to examine whether additional variance would be explained by one of the factors used in the first LME analysis. We found no significant difference in model fit between a model containing only Expectancy and a model containing Expectancy and Accuracy, $X^2(4) = 7.26$, $p = .123$. Likewise, we also found no difference in model fit between the Expectancy model and an Expectancy × Trial Type model, $X^2(4) = 2.40$, $p = .661$. Finally, a model containing Expectancy and Empathy did not account for significantly more variance than a model only containing Expectancy, $X^2(2) = 2.12$, $p = .346$.

# 4 | DISCUSSION

In this study we investigated the effects of empathy and expectancy on the processing of observed actions. As hypothesized, we found that the expectancy of the observed action was modulated by empathy in the sense that, depending on the empathy level of a participant, expectations developed differently during the course of the experiment so that toward the end of the experiment high empathy participants had stronger expectations concerning the accuracy of the observed response than low empathy participants. Subsequently, we found that the derived expectancy measures for each trial modulated observed response processing, as reflected in a frontocentral negative ERP component. Our analyses suggest that empathy affects observed response processing only indirectly, via its effect on expectation formation, as adding empathy to the statistical model did not explain significantly more variance of the analyzed ERP component than a model only containing expectancy. Similarly, the accuracy of the observed response and the fact whether the observed person

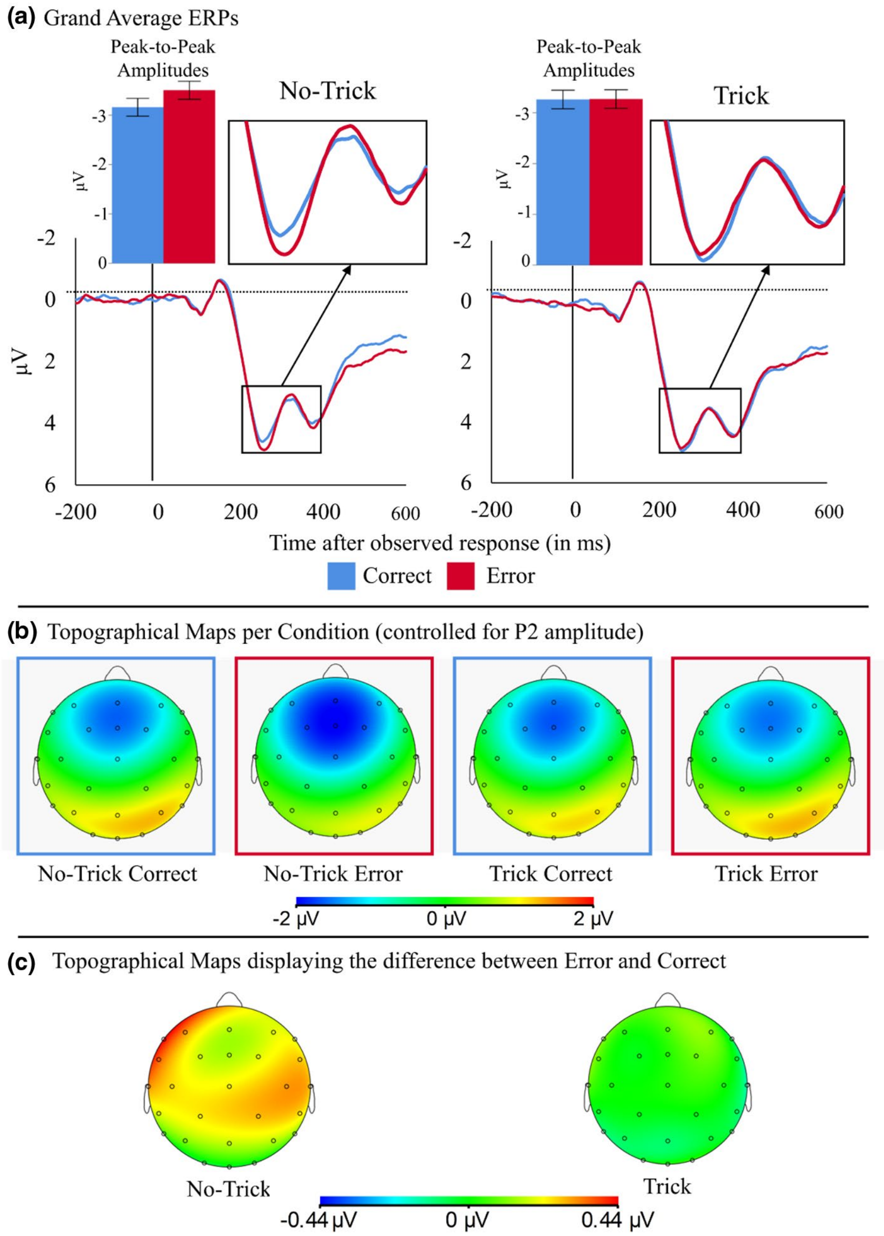



**FIGURE 3** Grand-average ERPs and topographies for observed error and correct responses in trick and no-trick trials. (a) Grand Average ERPs pooled over Fz, FCz, Cz, FC1 and FC1 after observed correct and error responses for trick and no-trick trials. The FRN complex (P2 followed by FRN) is highlighted. Additionally, mean peak-to-peak amplitudes for correct and error responses for trick and no-trick trials are displayed in bar plots. (b) Topographies of the highest negative peak at the pooled signal in all conditions relative to the preceding positive peak are shown (in accordance with the peak-to-peak measure used in the analyses). (c) Topographies of the difference of correct and error trials in both trial types at the highest negative peak in the error-correct difference wave at the pooled signal

had a true or false belief did not have significant effects on observed response processing.

## 4.1 | Expectancy formation and empathy

We measured participants' expectations by prompting, after a regular trial, which shell the participants believed the observed player to choose. Expectations were modulated by employing a true- and a false-belief condition that should lead to participants expecting correct answers in the true- and errors in the false-belief condition. We indeed found a respective effect of Trial Type in the prompt questions, indicating that the expectation modulation worked. This was in accordance with previous research employing this paradigm (Albrecht & Bellebaum, 2021; Bellebaum et al., 2020; Kobza & Bellebaum, 2013). In addition, we found that expectations of low empathy participants in the false-belief (trick) condition became less strong over the course of the experiment, while expectations did not change significantly during the

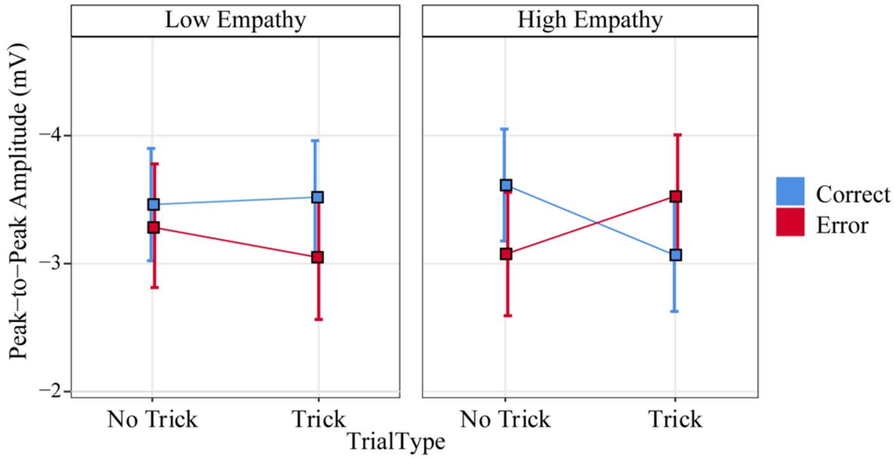



**FIGURE 4** Peak-to-peak amplitudes as a function of Empathy, Accuracy, and Trial Type, calculated in an LME analysis based on single trial ERPs of the frontocentral negative component. Error bars represent confidence intervals

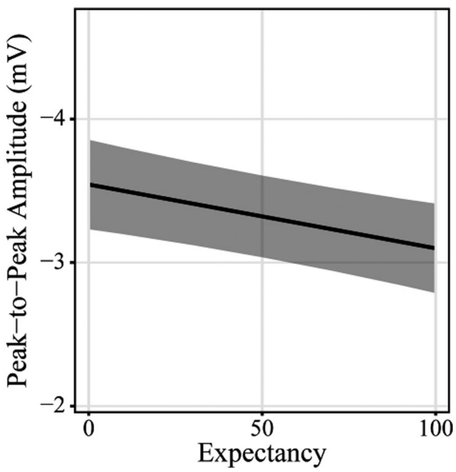



**FIGURE 5** Peak-to-peak amplitudes as a function of expectancy, calculated in an LME analysis based on single trial ERPs of the frontocentral negative component. Error measures represent confidence intervals

experiment for low empathy participants in the true-belief condition and for high empathy participants in both conditions. This led to differences in expectancies between high and low empathy participants, with more pronounced expectations in those with high empathy scores.

There are two possible explanations for the finding that expectations changed during the experiment exclusively for trick trials in low empathizers. Firstly, the induced expectancies (correct answers in no-trick, errors in trick trials) did not match the observed players' actual responses, which were correct in 50% of the trials in all conditions. During the course of the experiment, the participants might have noticed the discrepancy between their expectations at the beginning of the experiment and the actual answers and may have adjusted their expectancies. However, this seems unlikely, as such an effect would have appeared in both true- and false-belief trials and probably for both high and low empathy participants. Instead, the relatively stable expectations in most conditions match studies on the confirmation bias (Nickerson, 1998): When interpreting ambiguous events, participants tend to weigh events that support their initial beliefs more highly than vice versa (see Talluri et al., 2018; Urai et al., 2019).

We think that a second explanation for the change of expectations across the experiment is more likely, namely that for low empathy participants, it is difficult to uphold performance in trick-trials as these require the most empathy: previous research has shown that empathy is required to form expectations about others' actions in false-belief tasks (Birch & Bloom, 2007; Wellman et al., 2001). Less empathy should be required to form expectations in a true-belief condition. Several studies suggest that cognitive load may lead to reduced empathizing. Epley et al. (2004) found that participants were more likely to interpret situations from an egocentric perspective (rather than another person's) when under time pressure. Apperly et al. (2008) observed an effect of false-belief on processing costs when the false-belief was presented in the first of two sentences, but not when it was presented in the second, indicating a decrease in false-belief processing over time. Meyer et al. (2012) proposed the existence of what they called a social working memory system: while brain regions associated with mentalizing (temporoparietal junction, frontoparietal regions) were less active in previous studies with increasing cognitive or perceptual load, Meyer et al. found increased activation both in regions associated with working memory and regions associated with mentalizing when increasing social cognitive load. The authors argue that to be able to mentalize, specific information and assumptions about the other person have to be maintained and manipulated, which requires a form of working memory specified for social information. Meyer et al. indeed found lower task performance when "social load" was high. In addition, they found that activity in the medial prefrontal cortex and precuneus correlated with perspective taking abilities, which suggests that social working memory might be modulated by individual trait empathy. This explanation might also be strengthened by results we found in a previous study (Albrecht & Bellebaum, 2021) with a similar, but more complex, paradigm than in Bellebaum et al. (2020). In this paradigm that probably required more cognitive effort to form expectations, we found an effect of empathy also on averaged expectancies. In the current paradigm, forming expectations in the no-trick condition should be relatively straight-forward and less empathizing should be necessary (as the information given to the observed player and the observer participants is the same). In the trick condition, however, more empathizing should be necessary (as the observer participant has different knowledge compared with the player), which would mean an increased social load (see Meyer et al., 2012). Based on the social working memory system theory (Meyer & Lieberman, 2012; Meyer et al., 2012) it may be assumed that highly empathic participants, as they have greater social capacities, are able to empathize fairly well also in the trick condition, while for low empathic participants, this condition is socially and cognitively more demanding, and thus it might be harder for them to maintain a rather high social workload during the course of the experiment (Apperly et al.,2008; Epley et al., 2004; Meyer & Lieberman, 2012; Meyer et al., 2012). This effect would not show in the no-trick condition because social load should be lower in this condition.

In one previous study by our group, we found no effect of empathy on expectation formation (Bellebaum et al., 2020). However, the sample in this study showed a rather strong ceiling/floor effect regarding the expectancy of observed correct responses in the no-trick and trick trials, respectively, so that a possible modulation of empathy on expectancy would not be visible in these data. In a different variant of the paradigm with more variance in expectancy measures we recently found that empathy affected expectancy (Albrecht & Bellebaum, 2021). In addition, the study by Bellebaum et al. (2020) did not consider changes over the experimental course, and as an effect of empathy only emerged in interaction with trial progression across the experiment, possible effects of empathy on expectation formation may have been missed in our previous study.

## 4.2 | Frontocentral negative ERP component and expectancy

The first part of the analysis of the ERP data is a replication of the results by Bellebaum et al. (2020) with a larger, partially overlapping sample using single-trial data. This analysis already suggests that the frontocentral negative ERP component following observed responses is modulated by empathy and expectation, with variations in the latter being modulated via true- and false-belief conditions with the factor Trial Type. This shows that single-trial ERP analyses with LMEs yield comparable results as analyses based on average ERPs. Previous studies employing single-trial EEG analysis have employed a variety of approaches, including model-based regression analysis (Burnside et al., 2019; Pornpattananangkul et al., 2019) and machine learning (Stewart et al., 2014; Wirth et al., 2018). In few exploratory analyses, single-trial EEG data have been analyzed with LMEs: Frömer et al. (2018) combined cluster-based permutation tests and LME analyses. Spinnato et al. (2015) used LME to calculate classifiers in order to discriminate between errors and correct answers in EEG data. The present study contributes to this by introducing a rather simplified method to analyze single-trial EEG data that provides results comparable with more traditional analyses.

One aspect of the result pattern might appear inconclusive at first sight. In the first EEG analysis, we found effects of the experimental factors on ERP amplitudes in high empathy participants, but the behavioral data indicated that expectancies were affected in low empathy participants. However, in the behavioral data, this variation was dependent on the trial course, which was not considered in the EEG analysis. Moreover, the change in expectancies over time in low empathy participants led to differences in expectancies compared with high empathy participants. In the latter, expectancies were more pronounced which presumably led to more pronounced processing differences between conditions.

When Trial Type and Accuracy were replaced by single-trial expectancy values derived from individual statistical models for the development of expectancy over the course of the experiment, we found a main effect of the factor Expectancy, which indicated that when expectancy was low (observed response was unexpected), amplitudes were higher in contrast to when expectancy was high (response was expected). We found that Trial Type (true or false belief) did not account for significantly more variance when included in the model, suggesting that the differences found between the Trial Types in the first analysis was due to differences in expectancy and no other processes were at play. The same was true for Accuracy, which supports findings that it is expectancy, not valence, that influences ERPs that were previously linked to error monitoring (Desmet et al., 2014; Ferdinand et al., 2012; Jessup et al., 2010; Oliveira et al., 2007; Schiffer et al., 2014; Wessel et al., 2012, 2014; Zubarev & Parkkonen, 2018). We also found that a model including trait empathy in addition to expectancy did not explain significantly more variance than a model not including trait empathy. This is consistent with our hypothesis that empathy influences the frontocentral negative component only indirectly because it affects expectancy formation. The expectancies then influence observed response processing as reflected by the frontocentral negative component. As the behavioral data indicate, expectation formation (at least in the false-belief condition) is dependent on empathy. The findings are thus consistent with the results we obtained in a related recent study with a variant of the present paradigm. In that study we found that empathy affected expectations and observed response processing in parallel (Albrecht & Bellebaum, 2021). The results are, however, not directly comparable due to differences in the paradigm and, as a consequence of this, differences in the timing of the ERP component that was modulated. In the present study, taking the modulation by expectancy into account by integrating the measured expectancies into the model, it seems that the effect of empathy on ERPs does not exceed the effect empathy has on expectation formation. This may explain inconsistent findings for a modulation of ERP components by trait empathy (Brazil et al., 2011; Clawson et al., 2014; Lockwood et al., 2015; Newman-Norlund et al., 2009; Shane et al., 2009), as this seems to depend on the task and also on the context which might modulate expectation formation.

For example, Brazil et al. (2011) and Clawson et al. (2014) studied patients with psychopathy or autism, respectively, who observed another person performing a Flanker task. However, Brazil et al. presented the target arrow and the response, measuring the oERN, while Clawson et al. presented the target arrow, distractors and response, and added a feedback screen, measuring a component time-locked to observed feedback, the observer feedback-related negativity. This meant that participants in Clawson et al.'s study had two advantages: First, because they saw the distractors (exactly the same as the observed person), it was easier for them to take the observed person's perspective and thus to form expectancies concerning the accuracy of the answer. Second, they received direct feedback on whether the answer was correct or not, whereas participants in Brazil et al.'s study had to deduce this information themselves, which in the study by Clawson et al. meant a lower cognitive load that might have allowed a stronger focus on the observation itself (which is consistent with our behavioral results, see above), which again might have led to an easier formation of expectancies. These differences might serve as an explanation as to why Brazil et al. found differences between healthy controls and patients, but Clawson et al. did not.

Our interpretation of the results of the present study does not necessarily speak against the notion that the ACC is involved both in empathy (Singer & Lamm, 2009) and action monitoring (Ridderinkhof et al., 2004; Taylor et al., 2007; van Schie et al., 2004; see also Koban & Pourtois, 2014). The ACC, together with other brain regions, is believed to be a contributor to ERP components previously associated with error processing (see Gehring et al., 2012; Koban & Pourtois, 2014). If there is no direct modulation of action monitoring ERP components by empathy, the ACC could either independently contribute to both processes, or different parts of the ACC could contribute to empathy and action monitoring. Imaging techniques that have a better spatial resolution, such as fMRI, might be employed to further investigate this aspect. In any case, our data suggest that, for observed actions at least, empathy and expectancy are closely coupled and that the latter affects action monitoring.

## 4.3 | Limitations

While the experiment was constructed in a way that observer participants should have the impression that they observed a real person (by introducing the player by showing both a name and a picture and showing the player's hand during the trials), participants did not see, for example, a video of the player or even a “real” player. Empathy still seemed to be important for the current task, but it is possible that it might play even a greater role in more realistic settings.

We described the development of a binary variable (the answer to the prompt questions was either “right” or “left”) with a mathematical model in order to derive a continuous measure of expectancy for every trial, similar to what has been done for analyses of prediction errors before (e.g., see Burnside et al., 2019; Sutton & Barto, 1998). However, we did not measure expectancies for every trial, but only for 24 trials per Trial Type, distributed over the course of the experiment. We believe that these data were sufficient to calculate estimates for the expectancies of every participant in every trial, but it is likely that these models are not perfectly accurate and more data might be required to increase the models' accuracy. It is important to note that the models were linear, but expectations may have changed in a nonlinear way throughout the trials, which again would have resulted in less accurate models of the participants' expectations.

We also found that single-trial EEG data can be analyzed with mixed linear models in a way that the results are comparable to more traditional analyses. However, it should be noted that the method used in this study to extract single-trial data still considers only a fraction of the acquired EEG data. Extracting single-trial amplitudes at the time points of the average peaks allowed us to take into account variance of the data that would normally be lost due to averaging. In the current study, we aimed to apply analysis methods as similar as possible compared with previous studies on observed response processing that employed this paradigm (Bellebaum et al., 2020; Kobza & Bellebaum, 2013) or to other studies on action observation ERPs (e.g., Bates et al., 2005; Carp et al., 2009). However, a large part of the ERP data were still not considered, including data on other than the pooled electrodes and data on other than the determined latencies. Possible intertrial differences in the latency of the frontocentral negative component are also not considered.

## 4.4 | Conclusion

We found that the participants' expectation formation concerning observed responses and the development of expectations across the experiment was dependent on empathy: expectations became less strong over the course of the experiment only for low empathy participants in the false-belief condition, which led to more pronounced expectations in high than in low empathy participants. We demonstrated that these expectancies then shape ERP responses to observed actions, which is in accordance with previous findings by our group (Albrecht & Bellebaum, 2021; Bellebaum et al., 2020; Kobza & Bellebaum, 2013). By applying single-trial analysis and matching behavioral to EEG data, we showed that the influence of empathy on a negative frontocentral component was only indirect: empathy appears to influence expectancy formation, which then in turn affects the frontocentral negative component. The fact that the factor accuracy did not add significantly to the explanation of variance in the ERPs supports an existing body of literature indicating that it is expectations, not action valence, that primarily drive observed response processing. Future studies may investigate the interplay of empathy and expectancy further by using expectancy measures obtained in each trial.

## ACKNOWLEDGEMENT

The authors thank S. Kenneh, J. Ternig and M. Wollmer for their help in data acquisition. Open access funding enabled and organized by ProjektDEAL.

## CONFLICT OF INTEREST

The authors report no conflict of interest.

## AUTHOR CONTRIBUTIONS

**Christine Albrecht:** Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Visualization; Writing-original draft. **Christian Bellebaum:** Conceptualization; Methodology; Project administration; Resources; Supervision; Writing-review & editing.

## ORCID

*Christine Albrecht* https://orcid.org/0000-0002-8388-2533

## REFERENCES

Albrecht, C., & Bellebaum, C. (2021). Effects of trait empathy and expectation on the processing of observed actions. *Cognitive, Affective, & Behavioral Neuroscience*, *21*, 156–171. https://doi.org/10.3758/s13415-020-00857-7

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344. https://doi.org/10.1038/nn.2921

Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. *Cognition*, *106*(3), 1093–1108. https://doi.org/10.1016/j.cognition.2007.05.005

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163–175. https://doi.org/10.1023/B:JADD.0000022607.19833.00

Bates, A. T., Patel, T. P., & Liddle, P. F. (2005). External behavior monitoring mirrors internal behavior monitoring. *Journal of Psychophysiology*, *19*(4), 281–288. https://doi.org/10.1027/0269-8803.19.4.281

Bellebaum, C., Ghio, M., Wollmer, M., Weismüller, B., & Thoma, P. (2020). The role of trait empathy in the processing of observed actions in a false-belief task. *Social Cognitive and Affective Neuroscience*, *15*(1), 53–61. https://doi.org/10.1093/scan/nsaa009

Birch, S., & Bloom, P. (2007). The curse of knowledge in reasoning about false belief. *Psychological Science*, *18*, 382–386. https://doi.org/10.1111/j.1467-9280.2007.01909.x

Brazil, I. A., Mars, R. B., Bulten, B. H., Buitelaar, J. K., Verkes, R. J., & de Bruijn, E. R. A. (2011). A neurophysiological dissociation between monitoring one's own and others' actions in psychopathy. *Biological Psychiatry*, *69*(7), 693–699. https://doi.org/10.1016/j.biopsych.2010.11.013

Brown, E. C., & Brüne, M. (2012). The role of prediction in social neuroscience. *Frontiers in Human Neuroscience*, *6*, 147. https://doi.org/10.3389/fnhum.2012.00147

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, *107*(32), 14431–14436. https://doi.org/10.1073/pnas.1003111107

Burnside, R., Fischer, A. G., & Ullsperger, M. (2019). The feedback-related negativity indexes prediction error in active but not observational learning. *Psychophysiology*, *56*(9), e13389. https://doi.org/10.1111/psyp.13389

Carp, J., Halenar, M. J., Quandt, L. C., Sklar, A., & Compton, R. J. (2009). Perceived similarity and neural mirroring: Evidence from vicarious error processing. *Social Neuroscience*, *4*(1), 85–96. https://doi.org/10.1080/17470910802083167

Clawson, A., Clayson, P. E., Worsham, W., Johnston, O., South, M., & Larson, M. J. (2014). How about watching others? Observation of error-related feedback by others in autism spectrum disorders. *International Journal of Psychophysiology*, *92*(1), 26–34. https://doi.org/10.1016/j.ijpsycho.2014.01.009

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in, Psychology*, *10*, 85.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

de Bruijn, E. R. A., de Lange, F. P., von Cramon, D. Y., & Ullsperger, M. (2009). Where errors are rewarding. *The Journal of Neuroscience*, *29*(39), 12183–12186. https://doi.org/10.1523/JNEUROSCI.1751-09.2009

de Bruijn, E. R. A., & von Rhein, D. T. (2012). Is your error my concern? An event-related potential study on own and observed error detection in cooperation and competition. *Frontiers in Neuroscience*, *6*, 8. https://doi.org/10.3389/fnins.2012.00008 de Haen, J. (n.d.). *Deutsche version der Cambridge behavior scale*. http://docs.autismresearchcentre.com/tests/EQ_Deutsch.pdf

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological Science*, *5*(5), 303–305. https://www.jstor.org/stable/40063122

Desmet, C., Deschrijver, E., & Brass, M. (2014). How social is error observation? The neural mechanisms underlying the observation of human and machine errors. *Social Cognitive and Affective Neuroscience*, *9*(4), 427–435. https://doi.org/10.1093/scan/nst002

Donnarumma, F., Costantini, M., Ambrosini, E., Friston, K., & Pezzulo, G. (2017). Action perception as hypothesis testing. *Cortex*, *89*, 45–60. https://doi.org/10.1016/j.cortex.2017.01.016

Epley, N., Keysar, B., van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*(6), 447–455. https://doi.org/10.1016/0013-4694(91)90062-9

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: A tutorial. *Biological Psychology*, *51*(2–3), 87–107. https://doi.org/10.1016/s0301-0511(99)00031-9

Ferdinand, N. K., Mecklinger, A., Kray, J., & Gehring, W. J. (2012). The processing of unexpected positive response outcomes in the mediofrontal cortex. *The Journal of Neuroscience*, *32*(35), 12087–12092. https://doi.org/10.1523/JNEUROSCI.1410-12.2012

Ferguson, H. J., Cane, J. E., Douchkov, M., & Wright, D. (2015). Empathy predicts false belief reasoning ability: Evidence from the N400. *Social Cognitive and Affective Neuroscience*, *10*(6), 848–855. https://doi.org/10.1093/scan/nsu131

Flanagan, J. R., & Johansson, R. S. (2003). Action plans used in action observation. *Nature*, *424*(6950), 769–771. https://doi.org/10.1038/nature01861

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, *3*, 151. https://doi.org/10.3389/fpsyg.2012.00151

Frömer, R., Maier, M., & Abdel Rahman, R. (2018). Group-level EEG-processing pipeline for flexible single trial-based analyses including linear mixed models. *Frontiers in Neuroscience*, *12*, 48. https://doi.org/10.3389/fnins.2018.00048

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385–390. https://www.jstor.org/stable/40062567

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (Eds.). (2012). *The Oxford handbook of event-related potential. The error-related negativity (ERN/Ne)*. Oxford University Press.

Hajcak, G., Moser, J. S., Yeung, N., & Simons, R. F. (2005). On the ERN and the significance of errors. *Psychophysiology*, *42*(2), 151–160. https://doi.org/10.1111/j.1469-8986.2005.00270.x

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., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G. H., & Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*(5), 497–498. https://doi.org/10.1038/nn1238

Jessup, R. K., Busemeyer, J. R., & Brown, J. W. (2010). Error effects in anterior cingulate cortex reverse when error likelihood is high. *The Journal of Neuroscience*, *30*(9), 3467–3472. https://doi.org/10.1523/JNEUROSCI.4130-09.2010

Kang, S. K., Hirsh, J. B., & Chasteen, A. L. (2010). Your mistakes are mine: Self-other overlap predicts neural response to observed errors. *Journal of Experimental Social Psychology*, *46*(1), 229–232. https://doi.org/10.1016/j.jesp.2009.09.012

Koban, L., & Pourtois, G. (2014). Brain systems underlying the affective and social monitoring of actions: An integrative review. *Neuroscience and Biobehavioral Reviews*, *46*(Pt 1), 71–84. https://doi.org/10.1016/j.neubiorev.2014.02.014

Koban, L., Pourtois, G., Bediou, B., & Vuilleumier, P. (2012). Effects of social context and predictive relevance on action outcome monitoring. *Cognitive, Affective & Behavioral Neuroscience*, *12*(3), 460–478. https://doi.org/10.3758/s13415-012-0091-0

Koban, L., Pourtois, G., Vocat, R., & Vuilleumier, P. (2010). When your errors make me lose or win: Event-related potentials to observed errors of cooperators and competitors. *Social Neuroscience*, *5*, 360–374. https://doi.org/10.1080/17470911003651547

Kobza, S., & Bellebaum, C. (2013). Mediofrontal event-related potentials following observed actions reflect an action prediction error. *The European Journal of Neuroscience*, *37*(9), 1435–1440. https://doi.org/10.1111/ejn.12138

Koller, I., & Lamm, C. (2014). Item response model investigation of the (German) interpersonal reactivity index empathy questionnaire. *European Journal of Psychological Assessment*, *31*, 211–221. https://doi.org/10.1027/1015-5759/a000227

Lockwood, P. L., Apps, M. A. J., Roiser, J. P., & Viding, E. (2015). Encoding of vicarious reward prediction in anterior cingulate cortex and relationship with trait empathy. *The Journal of Neuroscience*, *35*(40), 13720–13727. https://doi.org/10.1523/JNEUROSCI.1703-15.2015

Marco-Pallarés, J., Krämer, U. M., Strehl, S., Schröder, A., & Münte, T. F. (2010). When decisions of others matter to me: An electrophysiological analysis. *BMC Neuroscience*, *11*, 86. https://doi.org/10.1186/1471-2202-11-86

Meyer, M. L., & Lieberman, M. D. (2012). Social working memory: Neurocognitive networks and directions for future research. *Frontiers in Psychology*, *3*, 571. https://doi.org/10.3389/fpsyg.2012.00571

Meyer, M. L., Spunt, R. P., Berkman, E. T., Taylor, S. E., & Lieberman, M. D. (2012). Evidence for social working memory from a parametric functional MRI study. *Proceedings of the National Academy of Sciences*, *109*(6), 1883–1888. https://doi.org/10.1073/pnas.1121077109

Miltner, W., Brauer, J., Hecht, H., Trippe, R., & Coles, M. (2004). Parallel brain activity for self-generated and observed errors. *Journal of Psychophysiology*, *18*(4), 205.

Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A. J., Schweizer, S., Frith, C. D., & Dalgleish, T. (2009). A key role for similarity in vicarious reward. *Science (New York, N.Y.)*, *324*(5929), 900. https://doi.org/10.1126/science.1170539

Newman-Norlund, R. D., Ganesh, S., van Schie, H. T., de Bruijn, E. R. A., & Bekkering, H. (2009). Self-identification and empathy modulate error-related brain activity during the observation of penalty shots between friend and foe. *Social Cognitive and Affective Neuroscience*, *4*(1), 10–22. https://doi.org/10.1093/scan/nsn028

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Ninomiya, T., Noritake, A., Ullsperger, M., & Isoda, M. (2018). Performance monitoring in the medial frontal cortex and related neural networks: From monitoring self actions to understanding others' actions. *Neuroscience Research*, *137*, 1–10. https://doi.org/10.1016/j.neures.2018.04.004

Oliveira, F. T. P., McDonald, J. J., & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error related: Expectancy deviation and the representation of action-outcome associations. *Journal of Cognitive Neuroscience*, *19*(12), 1994–2004. https://doi.org/10.1162/jocn.2007.19.12.1994

Paulus, C. (2009). *Der Saarbrücker Persönlichkeitsfragebogen SPF (IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index [The SPF (IRI) for the measure of empathy: Pschometric evaluation of the German interpersonal reactivity index]*. http://psydok.sulb.uni-saarland.de/volltexte/2009/2363

Pornpattananangkul, N., Grogans, S., Yu, R., & Nusslock, R. (2019). Single-trial EEG dissociates motivation and conflict processes during decision-making under risk. *NeuroImage*, *188*, 483–501. https://doi.org/10.1016/j.neuroimage.2018.12.029

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science (New York, N.Y.)*, *306*(5695), 443–447. https://doi.org/10.1126/science.1100301

Schiffer, A.-M., Krause, K. H., & Schubotz, R. I. (2014). Surprisingly correct: Unexpectedness of observed actions activates the medial prefrontal cortex. *Human Brain Mapping*, *35*(4), 1615–1629. https://doi.org/10.1002/hbm.22277

Shane, M. S., Stevens, M., Harenski, C. L., & Kiehl, K. A. (2008). Neural correlates of the processing of another's mistakes: A possible underpinning for social and observational learning. *NeuroImage*, *42*(1), 450–459. https://doi.org/10.1016/j.neuroimage.2007.12.067

Shane, M. S., Stevens, M. C., Harenski, C. L., & Kiehl, K. A. (2009). Double dissociation between perspective-taking and empathic-concern as predictors of hemodynamic response to another's mistakes. *Social Cognitive and Affective Neuroscience*, *4*(2), 111–118. https://doi.org/10.1093/scan/nsn043

Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences*, *1156*, 81–96. https://doi.org/10.1111/j.1749-6632.2009.04418.x

Spinnato, J., Roubaud, M.-C., Burle, B., & Torrésani, B. (2015). Detecting single-trial EEG evoked potential using a wavelet domain linear mixed model: Application to error potentials classification. *Journal of Neural Engineering*, *12*(3), 36013. https://doi.org/10.1088/1741-2560/12/3/036013

Stewart, A. X., Nuthmann, A., & Sanguinetti, G. (2014). Single-trial classification of EEG in a visual object task using ICA and machine learning. *Journal of Neuroscience Methods*, *228*, 1–14. https://doi.org/10.1016/j.jneumeth.2014.02.014

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

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology*, *28*(19), 3128–3135.e8. https://doi.org/10.1016/j.cub.2018.07.052

Taylor, S. F., Stern, E. R., & Gehring, W. J. (2007). Neural systems for error monitoring: Recent findings and theoretical perspectives. *The Neuroscientist*, *13*(2), 160–172. https://doi.org/10.1177/1073858406298184

Ullsperger, M., & Cramon, D. (2004). Neuroimaging of performance monitoring: Error detection and beyond. *Cortex*, *40*, 593–604. https://doi.org/10.1016/S0010-9452(08)70155-2

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

Urai, A. E., de Gee, J. W., Tsetsos, K., & Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. *eLife*, *8*, e46331. https://doi.org/10.7554/eLife.46331

van Schie, H. T., Mars, R. B., Coles, M. G. H., & Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, *7*(5), 549–554. https://doi.org/10.1038/nn1239

Wang, L., Tang, D., Zhao, Y., Hitchman, G., Wu, S., Tan, J., & Chen, A. (2015). Disentangling the impacts of outcome valence and outcome frequency on the post-error slowing. *Scientific Reports*, *5*, 8708. https://doi.org/10.1038/srep08708

Weller, L., Schwarz, K. A., Kunde, W., & Pfister, R. (2018). My mistake? Enhanced error processing for commanded compared to passively observed actions. *Psychophysiology*, *55*(6), e13057. https://doi.org/10.1111/psyp.13057

Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*, 655–684. https://doi.org/10.1111/1467-8624.00304

Wessel, J. R., Danielmeier, C., Morton, J. B., & Ullsperger, M. (2012). Surprise and error: Common neuronal architecture for the processing of errors and novelty. *The Journal of Neuroscience*, *32*(22), 7528–7537. https://doi.org/10.1523/JNEUROSCI.6352-11.2012

Wessel, J. R., Klein, T. A., Ott, D. V. M., & Ullsperger, M. (2014). Lesions to the prefrontal performance-monitoring network disrupt neural processing and adaptive behaviors after both errors and novelty. *Cortex*, *50*, 45–54. https://doi.org/10.1016/j.cortex.2013.09.002

Wirth, C., Lacey, E., Dockree, P., & Arvaneh, M. (2018). Single-trial EEG classification of similar errors. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society*, *2018*, 1919–1922. https://doi.org/10.1109/EMBC.2018.8512700

Zubarev, I., & Parkkonen, L. (2018). Evidence for a general performance-monitoring system in the human brain. *Human Brain Mapping*, *39*(11), 4322–4333. https://doi.org/10.1002/hbm.24273


## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

**TABLE S1** Demographic and behavioral data for women and men plus additional inferential statistics comparing the groups

**FIGURE S1** Histogram of Empathy (EQ) values in the sample

**FIGURE S2** Histogram of Expectancies in the trick and no-trick condition for the initial (women-only) and the complete sample as well as direct comparisons

**FIGURE S3** Behavioral expectancy data as a function of Trial Type, Trial Number and Empathy (resolved for empathy)

**FIGURE S4** Behavioral expectancy data as a function of Trial Type and Empathy
**TABLE S2** Statistical data for behavioral LME analysis
**TABLE S3** Statistical data for ERP LME analysis for Accuracy × Trial Type × Empathy
**FIGURE S5** Scatterplot of the peak-to-peak amplitude of the frontocentral negative component as a function of Empathy, Accuracy and Trial Type including the regression lines for each condition
**TABLE S4** Statistical data for ERP LME analysis for expectancy
**FIGURE S6** Scatterplot of the peak-to-peak amplitude of the frontocentral negative component as a function of expectancy including the regression line

**How to cite this article:** Albrecht, C., & Bellebaum, C. (2021). Disentangling effects of expectancy, accuracy, and empathy on the processing of observed actions. *Psychophysiology*, 58, e13883. https://doi.org/10.1111/psyp.13883

# Slip or Fallacy? Effects of Error Severity on Own and Observed Pitch Error Processing in Pianists

Christine Albrecht[1*] and Christian Bellebaum[1]

[1] Institute of Experimental Psychology, Heinrich Heine University Düsseldorf, Germany

* corresponding author

Christine Albrecht

Institute of Experimental Psychology

Heinrich Heine University Düsseldorf

Universitätsstraße 1

building 23.03

room number 00.89

40225 Düsseldorf

Germany

christine.albrecht@hhu.de

**Abstract**

Errors elicit a negative mediofrontal event-related potential (ERP), for both own errors (error-related negativity; ERN) and observed errors (here referred to as observer mediofrontal negativity; oMN). It is yet unclear, however, if the action monitoring system codes action valence as an all-or-nothing phenomenon or if the system differentiates between errors of different severity. We investigated this question by recording electroencephalography (EEG) data of pianists playing themselves (Experiment 1) or watching others playing (Experiment 2). Piano pieces designed to elicit large errors were used. While active participants' ERN amplitudes differed between small and large errors, observers' oMN amplitudes did not. The different pattern in the two groups of participants was confirmed in an exploratory analysis comparing ERN and oMN directly. Additionally, only ERN amplitudes were sensitive for later error correction. Post-hoc analyses suggested that the oMN is more strongly driven by the expectancy of the eliciting event than the ERN. We suspect that both prediction and action mismatches can be coded in action monitoring systems, depending on the task, and a need-to-adapt signal is sent whenever mismatches happen to indicate the magnitude of the needed adaptation.

*Keywords:* Error Severity, Action Monitoring, Observed Action Monitoring, ERN, oERN

**Slip or Fallacy? Effects of Error Severity on Own and Observed Pitch Error Processing in Pianists**

Over the last 30 years, researchers have investigated the neural correlates of error processing mostly by treating errors as an all-or-nothing phenomenon (either error or correct; e.g. Falkenstein et al., 1991; Gehring et al., 1993; Jessup et al., 2010; Ullsperger et al., 2014). Contrasting errors vs. correct actions showed that error processing involves several areas at the medial wall of the prefrontal cortex (this region will be referred to as medial prefrontal cortex or mPFC), including the anterior cingulate cortex (ACC; Debener et al., 2005; Ullsperger et al., 2014). An event-related potential (ERP) component investigated in the context of error processing is the error-related negativity (ERN), a negative-going frontocentral deflection that peaks around 100 ms after an erroneous response (Falkenstein et al., 1991; Falkenstein et al., 2000; Gehring et al., 1993, for reviews, see Gehring et al., 2012; Holroyd & Coles, 2002). The ERN appears to be generated in mPFC, probably the ACC (Debener et al., 2005; Dehaene et al., 1994; Ridderinkhof et al., 2004; Taylor et al., 2007).

Not all researchers agree that the mPFC is primarily involved in error processing, and thus it has also been questioned whether the ERN reflects error processing per se. Apart from a conception in terms of conflict monitoring (Botvinick et al., 2001; Carter et al., 1998; Yeung et al., 2004), the more recent predicted-response outcome model (PRO model; Alexander & Brown, 2011), states that the mPFC activity reflects unexpected outcomes and actions rather than errors (see Gawlowska et al., 2018; Jessup et al., 2010; Wessel et al., 2012). Although there is initial evidence supporting this view (Jessup et al., 2010; Wessel et al., 2012), there is also reason to believe that the mPFC codes information that is particularly relevant for error processing.

In this context it is interesting to note that all of the above-mentioned findings neglect the fact that action valence can vary more gradually than just distinguishing right vs. wrong. In sports, music or many other motor-cognitive tasks, people can diverge from the correct movement on a scale from “perfect” to “completely wrong”. In everyday language, we use terms such as *slip* or *fallacy*, which also suggests that we distinguish between errors of different severity. The ACC receives input from both motor and cognitive brain areas and is supposedly involved in the planning and regulation of behavior (see Devinsky et al., 1995; Holroyd & Coles, 2002), making it a crossroad for correcting and adapting behavior. For this function, the system needs to know how much adaptation is needed - for example when a pianist hits a key one or two notes amiss and must adapt their hand position within milliseconds to hit the next note. Taking into account the function of the ACC, the variety of errors in everyday life, and the early processing needed for error (severity) detection in order to adapt behavior, it is conceivable that error severity is processed early after error commission in the time window of the ERN. We thus assume that the ERN as a fast indicator of information related to error processing codes action valence on a spectrum and not as an all-or-nothing phenomenon, thus reflecting error severity. This would also support the view that the mPFC/ACC is, at least partially, involved in representing performance accuracy and not entirely driven by event expectancy, as stated by the PRO model (Alexander & Brown, 2011, see for example Maier & Steinhauser, 2016 for conflicting results regarding the model).

Initial evidence supporting the assumption of a continuous encoding of error severity stems from studies comparing different types of responses yielding different error types (under-reach vs. over-reach, Murata & Katayama, 2005; hand vs. finger, Falkenstein et al., 2000; corrected vs. uncorrected, Paas et al., 2021). An effect of error size has been described in two paradigms in which wrong actions in either one (single error) or two (double error) dimensions were possible (Bernstein et al., 1995; Maier et al., 2008; Maier et al., 2012; Maier & Steinhauser, 2016): double errors led to significantly larger ERN amplitudes than single errors. These results, however, may also be explained by two parallel action monitoring processes for both dimensions, each coding accuracy in a binary fashion, that add up to an increased ERN. It has yet to be investigated whether different degrees of deviations from the aspired action indeed lead to correspondingly increased neural responses in action monitoring regions.

Interestingly, the processing of observed actions involves similar brain areas as the processing of self-actions, such as the mPFC, specifically the ACC (Yoshida et al., 2012, for a review, see Koban & Pourtois, 2014) and pre-supplementary and supplementary motor areas (Scangos et al., 2013), with additional activity, inter alia, in the superior temporal sulcus (Ninomiya et al., 2018), inferior frontal gyrus (Shane et al., 2008), and anterior insula (Cracco et al., 2016, for a review, see Koban & Pourtois, 2014). Accordingly, observed errors have been reported to elicit an ERP component corresponding to the ERN, the observer error-related negativity (oERN) at frontocentral sites (Bates et al., 2005; de Bruijn & Rhein, 2012; Miltner et al., 2004; van Schie et al., 2004). Source localization suggests the origin of the oERN also in the mPFC (van Schie et al., 2004), probably in the ACC (Miltner et al., 2004). Compared to the ERN, the oERN displays smaller amplitudes and peaks later relative to the eliciting event, which is an observed action and thus a visual stimulus rather than an own motor response, with the latency depending on the task (Bates et al., 2005; de Bruijn & Rhein, 2012; van Schie et al., 2004). Research in observed error processing, as in own error processing, has mostly focused on binary response classifications in terms of accuracy (e.g. Bates et al., 2005; de Bruijn & Rhein, 2012; Kobza & Bellebaum, 2013). Recent evidence from our lab indicated, however, that observed responses are processed primarily based on their expectancy and not their accuracy (Albrecht & Bellebaum, 2021a, 2021b), which might also lead to differences compared to active responding with respect to effects of error severity.

To study effects of error severity so-called sequential tasks such as typing or playing the piano (Herrojo Ruiz et al., 2009; Kalfaoğlu et al., 2018; Maidhof et al., 2009; Maidhof et al., 2013; Paas et al., 2021) appear to be a particularly suitable. In these tasks, errors are frequent and participants stay seated while performing a (highly-practiced) everyday motor task that is ecologically valid and not dependent on feedback (Herrojo Ruiz et al., 2009). Typically, the ERN occurs 20-100 ms before the response in sequential tasks (Herrojo Ruiz et al., 2009; Kalfaoğlu et al., 2018; Maidhof et al., 2009; Paas et al., 2021) and thus earlier than in tasks involving a single response (Falkenstein et al., 1991; Gehring et al., 1993). Maidhof et al. (2013) showed that potential errors are noticed earlier with regard to the registered keypress (probably due to earlier movement onset compared to non-sequential tasks), and earlier error registration results in shorter ERN latencies (Di Gregorio et al., 2022). Further, error monitoring and error severity processing are especially important for adaptation during sequential tasks.

In the present study, we thus conducted two experiments with pianists. In Experiment 1, participants played piano pieces which included frequent changes of hand positions, thereby provoking small and large errors. While participants played, both EEG and behavioral data were assessed. Videos recorded during Experiment 1 served as stimuli for Experiment 2, in which participants watched videos of other pianists performing while EEG data were assessed in the observers. With these experiments we aimed to investigate two main questions: First, are ERN amplitudes enhanced for larger compared to smaller errors? And second, is a similar effect found also for observed errors?

## Experiment 1

In Experiment 1 we studied effects of error severity on error processing during active piano playing. Apart from the neural processing of errors, the piano playing paradigm allows to investigate also relevant behavioral variables: First, post-event reaction times can be assessed. A relative slowing of reaction times after errors is a well-studied phenomenon (Rabbitt, 1966, 1969), possibly linked to an attentional shift towards the error (or unexpected event), resulting in an attention reorienting process back to the task that underlies the longer reaction times (Notebaert et al., 2009; Núñez Castellar et al., 2010). Post-error slowing is presumably modulated by activity in the ACC (Danielmeier et al., 2011; Debener et al., 2005; Fu et al., 2019), but findings on the relationship between ERN and post-error slowing are mixed (Chang et al., 2014; Debener et al., 2005; Gehring et al., 1993; Hajcak et al., 2003). Possibly, some factors influence post-error slowing and the ERN differently (such as expertise, Jentzsch et al., 2014, or error awareness, Nieuwenhuis et al., 2001), leading to a dissociation in respective tasks. Post-error slowing has also been observed in piano-playing tasks (Herrojo Ruiz et al., 2009; Paas et al., 2021). A second variable of interest is keypress volume (assessed as velocity), as error notes were played significantly more quietly than correct notes in previous piano playing studies (Herrojo Ruiz et al., 2009; Maidhof et al., 2009; Maidhof et al., 2013; Paas et al., 2021). As larger errors might lead to a larger focus of attention on the error, enhanced post-error slowing was expected for large compared to small errors. Additionally, quieter keypress volumes after errors compared to correct keypresses were expected, but as the processes behind the volume reduction are not yet established, we refrain from predicting differences between small and large errors regarding volume.

Person-related factors may also affect error processing, especially in musicians, who form a specific group of participants. First, musical training in children seems to enhance trait empathy (Hietolahti-Ansten & Kalliopuska, 1990; Kalliopuska & Ruókonen, 1986; Rabinowitch et al., 2012) and adult musicians show significantly higher empathy levels compared to non-musicians, with increased connectivity of the insula with, inter alia, the ACC (Gujing et al., 2019). Some studies suggest a positive effect of trait empathy on ERN amplitude (Larson et al., 2010; Santesso & Segalowitz, 2009). Although the findings are inconsistent (Amiruddin et al., 2017), potential effects of empathy should be considered when analyzing a high-empathy sample such as musicians.

Secondly, the musicians acquired in the present study are an expert sample, and expertise might alter action monitoring processes (Jentzsch et al., 2014; Rachaveti et al., 2020). For example, expert pianists rely little on external auditory feedback for detecting pitch errors (Finney, 1997; Finney & Palmer, 2003; Herrojo Ruiz et al., 2009; Maidhof et al., 2009). The altered action monitoring processes in experts might generate to other tasks outside their area of expertise: Jentzsch et al. (2014) found enhanced ERN amplitudes for highly compared to barely trained musicians in a Stroop task. Also, expert musicians exhibit less post-error slowing than non-musicians, at least in non-music tasks (Jentzsch et al., 2014). More experienced pianists thus might exhibit altered error processing compared to persons with less or no piano experience due to their specific training.

The paradigm also offered the opportunity to investigate additional factors possibly affecting action monitoring. First, we aimed to replicate findings by Paas et al. (2021) who found altered behavioral and electrophysiological responses to later corrected vs. uncorrected piano playing errors. The authors investigated only notes for which the next keypress was correct, and since the hand had to be adjusted to achieve this, corrected and uncorrected errors both must have been recognized by the player, so that differences in error recognition are unlikely to account for the finding. However, as error correction should be avoided in piano playing, later corrected errors probably lead attention away from the task and towards the error itself - later corrected errors thus might be perceived as subjectively more important. Indeed, effects of the subjective importance of errors on error processing have been observed (Ganushchak & Schiller, 2008; Hajcak et al., 2005). These findings constitute a challenge for the PRO model (Alexander & Brown, 2011), as does the finding that double errors elicit higher ERN amplitudes than single errors even when expectancies are accounted for (Maier & Steinhauser, 2016). Based on these findings, we hypothesized that the processing of active responses is not entirely driven by expectancy and that therefore participants show more pronounced ERNs for large compared to small errors (and for corrected than uncorrected errors), while all types of errors elicit an ERN relative to correct responses. To explore potential effects of expectancy, we took into account additional behavioral measures indirectly related to expectancies of the shown response as potential predictors of neural action monitoring processes: the difficulty of the respective note (mean error rate of each note by participant), the overall frequency per participant of each error type, and the insecurity with which the respective key was played (deviance from the mean volume of each participant). As suggested by Maier and Steinhauser (2016), we expected error severity to predict error processing beyond the effect of expectancy.

## Method

### Participants

We recruited experienced pianists to take part in the study via social media, person-to-person recruiting and flyers distributed at the university, music conservatory, and music schools.

Because the pieces included large steps between keys to induce errors and the pieces were thus difficult to learn, we suggested a minimum experience of 1500 hours spent with the instrument, although participants were allowed to take part with less experience if they were able to play the pieces fluently. Data from 25 participants were recorded. Of these, one had to be excluded due to technical difficulties during data acquisition and another 3 were excluded because they made less than 10 large errors. The remaining sample of 21 participants consisted of 12 cis-gender women and 9 cis-gender men between 17 and 34 years old (Mean [*M*] = 23.1 years, Standard Deviation [*SD*] = 4.2 years). Twenty of them were right-handed, one person was left-handed. All participants reported no previous neurological or psychiatric illnesses and no intake of medication that affected the nervous system. All participants took part voluntarily. The study is in compliance with the declaration of Helsinki and was approved by the ethics committee of the Faculty of Mathematics and Natural Sciences at Heinrich-Heine-University, Düsseldorf.

**Material**

We designed six pieces to be played with only the right hand. All pieces consisted of 96 sixteenth notes in 6 bars and ended with a seventh bar that consisted of a single whole note. To keep the physical distance between played keys constant, all pieces were written in C major and thus only played on white keys. The pieces kept to a general harmonic structure and the highest notes played could be interpreted as a melody, the remaining notes as accompaniment. The pieces were designed to require large hand movements to induce errors. The lowest key throughout the pieces was E3, the highest key was A5. Consecutive notes could differ between 1 and 10 white keys; the average difference was 4.98 white keys (*SD* = 1.88 keys). The pieces were written in MuseScore 3 (version 3.6.2, MuseScore BVBA, 2021). They are included in the supplementary material (Figure S1).

An automatically created recording was generated for each of the pieces (created with MuseScore 3, version 3.6.2, MuseScore BVBA, 2021) in which the melody parts of the pieces were pronounced. In the recording, pieces were played at 60 beats per minute (one beat = one quarter note), and tempo at the top of the score notation was also stated as 60 quarter notes per minute.

**Experimental Task and Setup**

The pieces as well as the recording were sent to each participant two weeks before testing. Participants were instructed to study the pieces in the next 14 days. They were told that they should be able to play the pieces with the right hand quite fluently, but that they should not strive for perfect sound and that occasional errors during play were acceptable. Participants were also told to practice in whatever tempo they felt comfortable. They were given an instruction to practice approximately 15 minutes a day (distributed as they saw fit). According to self-reports, the participants practiced the pieces 204.1 minutes on average (*SD* = 188.9 minutes, 45-840 minutes).

For data acquisition during the experiment, participants used a digital piano (Casio LK-S450 for most participants, two participants used a Yamaha YDP-144 R Arius). During the experiment, the keyboard was set on mute, so that participants could not hear themselves play. The piano was positioned in front of a desktop monitor (1920 x 1080 px) that served for visual stimulation. Participants could navigate through the experiment with their left hand and the lowest note on the keyboard. A Logitech BRIO webcam was connected to an additional laptop for recording the participants' hand from above during play for the videos used in Experiment 2. A picture of the setup can be seen in Figure 1. We recorded the Musical Instrument Digital Interface (MIDI) information of the played segments on the experiment computer. MIDI refers to the signal used by digital instruments to generate and communicate tones including note on- and offset, key and velocity (in piano playing, this corresponds to volume). Stimulus presentation, EEG trigger timing and MIDI recording was controlled with Python 3.7.5 using the packages psychopy (version 3.2.3; Peirce et al., 2019) and mido (version 1.2.9, Ole Martin Bjørndalen 2021, mido.readthedocs.io).

The experiment consisted of 60 sequences in total, 10 for each piece. Each sequence started with a score notation preview of the piece that was to be played (a picture of the first two bars, i.e. the first line, of the respective piece score notation, including the piece number). Participants could then start the recording which began with 4 metronome beats (1000 hz beeps) accompanied by the numbers 1 to 4 displayed on the screen. Subsequently, the score notation of the whole piece was displayed on the screen to allow participants to play from sheet. After they were done with playing the piece, participants ended the recording and proceeded to the next sequence. A display of the sequence structure can be seen in Figure 2.

Before the experiment, participants were asked in what tempo they had practiced the pieces. Accordingly, the metronome beats were set for each participant individually to a tempo slightly faster than the tempo in which they had practiced to increase difficulty. Participants were instructed to start playing right after the last metronome beat had been presented. They were further asked to keep to one tempo (loosely that of the metronome) during each sequence and to put emphasis on playing fluently, even if that meant making errors.

The 60 sequences were preceded by 3 practice sequences in which participants could get to know the procedure of a sequence, but in which they were shown a mock preview and no actual score notation during play. They were instructed to get familiar with the instrument and the procedure during these practice sequences, and to play whatever came to their mind.

## Assessment of Empathy and Expertise

### *Empathy*

We chose the empathy quotient (EQ) measured with the German version of the Cambridge Behavior Scale (Baron-Cohen & Wheelwright, 2004; de Haen, n.d.) to measure trait empathy. In the scale, participants are asked to rate their agreement to 60 statements on a 4-point Likert scale (from "strongly agree" to "strongly disagree"). The statements contain 20 distractor items. For positively scaled true items, participants receive 2 points for "strongly agree", 1 point for "slightly agree" and 0 points for "slightly disagree"/"strongly disagree". Scoring is reversed for negatively scaled true items. All points from true items are added up for the EQ sum score (0 to 80 points).

We additionally administered the German short version (Paulus, 2009) of the Interpersonal Reactivity Index (IRI; Davis, 1980, 1983). The IRI is another empathy measure containing subscales for cognitive and affective empathy. The index was included to enable post-hoc analyses as to which aspect of empathy contributed to a possible general empathy effect.

### *Piano Playing Expertise*

Expertise was defined as total hours spent with the instrument, calculated by multiplying the self-reported number of years of piano experience with the self-reported average hours of practice per week times 52 (number of weeks per year).

## EEG Recording

We recorded EEG signals at a 1000 Hz sampling rate with a 32-channel actiCap electrode cap (ActiCAP; Brain Products GmbH, Germany) with the software Brain Vision Recorder (version 1.20, Brain Products, Munich, Germany). The active silver/silver-chloride electrodes were attached according to the 10-20 system on 29 scalp sites, i.e. FCz (which was used as online reference), F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO9, P1, Pz, P2 and PO10. Additionally, we recorded the signal from both mastoids to use as offline reference. The ground electrode was placed at AFz. For electrooculogram (EOG) data, we placed two horizontal EOG (hEOG) electrodes at F9 and F10, respectively, and two vertical EOG (vEOG) electrodes at Fp2 and below the right eye. All impedances were kept below 10 kΩ.

An EEG marker was sent every $5^{th}$ keystroke to avoid a possible overlap of markers (Maidhof et al., 2009). The MIDI data allowed offline determination of markers for the remaining keystrokes. We conducted a pilot test for a possible delay between key press and marker by using a Tektronix TDS 210 oscilloscope. Key presses are transformed to audio signals by the digital instrument in real-time. In the test, we therefore compared onset times between the audio and marker signal. The markers were sent consistently 1.6 ms before tone onset across all tests.

**Procedure**

Participants received the piano pieces two weeks before the actual experiment in the lab. After arrival, participants gave informed written consent to take part in the study and completed a demographic questionnaire and empathy and expertise self-report measures.

Subsequently, EEG electrodes were attached to the scalp and participants started the experiment. Participants received written instructions and the experimenters were present during three practice sequences for questions and further explanations. At the start of the experiment, recordings of video, MIDI and EEG were started. The experiment lasted between 35 and 75 minutes, depending on the speed in which participants played. After completion of the experiment, participants received compensation in the form of either course credit or 40 €.

## Data Analyses

### *Behavioural Data Processing*

**Data Preprocessing and Definition of Event Types.** All following analysis steps were performed in MATLAB, version R2017b (Mathworks, Natick, Massachusetts, USA). We employed the MATLAB MIDI Toolbox (Eerola & Toiviainen, 2004) and a dynamic score matcher algorithm created by Large (1993; see also Palmer & van de Sande, 1993; Rankin et al., 2009) to compare the recorded MIDI signal with the correct score notation the participants had been asked to play. This procedure was used to determine the different types of trials for which ERP and behavioral data were compared (see below). The algorithm finds a so-called optimal match between two MIDI sequences and assigns every played note an attribute: match, substitution (a score notation note was replaced in the performance), addition (there was an added note in the performance that could not be matched to any notation note) and miss. All substitution events were defined as “uncorrected” errors (see also below).

We then calculated the interval in white keys for substitution events (and addition events, see post-hoc-analyses) between the correct score notation note and the corresponding performance note. Black keypresses were not considered in the analysis.

In the analyses, we included the event types correct, small error (one-note errors that were not corrected), and large error (two-note errors that were not corrected). All errors larger than two-not errors were excluded. Moreover, we only included error and correct events that were preceded and followed by a correctly played note, which also excludes correct notes played before or after miss events. Each of the 97 notes included in the score notation of each piece was played 10 times in the course of the experiment, which allowed us to calculate the note accuracy for every note as the percentage of times the note was played correctly. Only notes that had an accuracy higher or equal to 40% were considered in the analysis, to exclude notes that were played systematically wrong. Additionally, we only included notes for which at least one error trial and one correct trial was included, to avoid confounds of note selection.

For the purpose of exploratory analyses, we determined an additional event type, corrected errors (see Paas et al., 2021). Corrections barely occurred for large errors, which is why we concentrated on the comparison between small corrected and small uncorrected errors. Corrected errors were always marked as addition events, but addition events did not always represent subsequently corrected errors. Sometimes, participants seemed to have slipped *after* playing a note (pressed the same or an adjacent key). To select only corrected notes, we included addition notes that were closer to the following than to the previous note; other addition notes were marked as postslips. For corrected error events, we defined the following correct note in the score notation as the corresponding note to the performance note.

**Behavioral Dependent Variables.** Two behavioral measures served as dependent variables, which possibly differed between event types (correct, small error, large error) or between corrected and uncorrected errors. To investigate potential behavioral effects of error severity, namely on keypress volume and post-event slowing, the behavioral dependent variables Volume and Inter-Keypress-Interval (IKI) were assessed. Volume was defined as the recorded velocity in the MIDI signal of each note. IKI was defined as the difference between note onset time of the current and of the following note (see Paas et al., 2021). This maps the time delay between the event (correct, small or large error) and the subsequent correct keypress and serves as a measure of post-event reaction time, which is used to calculate post-error-slowing.

**Behavioral Variables as Potential Predictors of Action Monitoring**. Finally, we calculated three variables based on the behavioral data that could be related to participants' expectancy regarding the action outcome and that were used as potential predictors of the dependent behavioral and ERP measures reflecting action monitoring. First, the inverted note accuracy (the non-inverted note accuracy was previously used as inclusion criterion), i.e. the note error rate, was defined as a measure for the difficulty of any included note. The higher the error rate, the more difficult the note should be for the respective participant, and the more errors they should expect (see Albrecht & Bellebaum, 2021b, for a modulation of expectancy by difficulty). Second, the distribution of event types was different for participants depending on their performance, and we expected that the more frequent an event type was, the more it would be expected. The percentage of all event types (correct, small error, and large error) for each participant formed the variable Event Type Frequency. Third, piano players often play either more loudly or more quietly in passages in which they are insecure, so changes in velocity can serve as an indicator of (in)security. The less secure participants are about a passage, the more they should expect to make errors, which is why insecurity might influence expectancies. Insecurity was calculated as the absolute difference between the velocity of each played note and the mean velocity for each participant. All continuous measures that were considered subsequently in any analysis were scaled to lie between -0.5 and 0. 5 and then mean-centered.

***Behavioral Data Statistical Analysis***

For all statistical analyses, if not stated differently, we conducted single-trial linear mixed models (LME) analyses in R (version 3.5.3) using the package lme4 (version 1.1-23). According to best practice (Meteyard & Davies, 2020), all models should include all within-subject main and interaction effects as random effects, if this is possible without leading to model fit errors. For all subsequently described analyses, we performed an iterative process: all within-subject main and interaction effects were first included as random factors. If this led to model fit errors (singular fit or overfitting), we tested which random effect led to this error and removed this from the model. As most of our models included only the main effect Event Type, for some models this factor is included as random effect factor and for others not, depending on the model fit.

All trials of the respective statistical models (see below) for which the behavioral dependent measures IKI or volume value differed by more than two SDs from the mean values per participant and Event Type were removed for further analysis. On average, 4.6 % of trials (*SD* = 0.8 %, Maximum = 6.0 %) were excluded for each participant and condition for the IKI analysis, and 4.0 % of trials (*SD* = 0.72 %, Maximum = 5.5 %) were excluded for the volume analysis.

We conducted LME analyses, calculating separate models for dependent variables IKI (post-event reaction time) and volume (velocity). As independent variable, we set the three-level factor Event Type (correct, small error, large error). Small error was set as baseline condition to determine both the difference between correct and (small) errors (to analyze whether an ERN occurred) and between small and large errors. Consequently, we created the design matrix depicted in Table 1 based on simple coding. We included random intercepts and slopes for Event Type per participant into each model.

With Cook's Distance outlier detection (using the "influence" function of the package stats, version 4.02, in R) based on the calculated models (with a cut-off value of 4/(n-number of predictors-1)), we removed 3 participants from the IKI analysis (remaining *n* = 18, 17 – 34 years, *M* = 23.0 years, *SD* = 4.4 years, 9 women, 9 men) and 1 participant from the volume analysis (remaining *n* = 20, 17 – 34 years, *M* = 23.3 years, *SD* = 4.2 years, 11 women, 9 men). Subsequently, the models were recalculated with the new sample.

We also investigated possible modulations of IKI and volume by Empathy and Expertise (in interaction with Event Type). Model comparisons (on a Bonferroni-corrected Alpha level of .025; with lmerTest, version 3.1-3) were used to check if any model that included Empathy or Expertise explained significantly more variance than the original model. If one or both did, we checked whether adding both at the same time explained more variance than adding only one. To investigate the simplest possible model, we only included Empathy or Expertise if one or both explained significantly more variance, and calculated further statistical results with all included predictors on the thus determined best model.

***EEG Data Preprocessing***

We recoded the EEG marker files offline by synchronizing the markers sent every five notes with the recorded MIDI data using MATLAB. The new markers were then written into new marker files which were loaded into Brain Vision Analyzer (Brain Products, Munich, Germany). Subsequently, we applied a 0.5 Hz high-pass and 30 Hz low-pass filter to the data (as suggested by Luck, 2014). As participants read score notations while they played and were not prevented from looking down on their hand (both to obtain maximum ecological validity), vertical and horizontal eye movements occurred frequently during the experiment and the corresponding artefacts had to be removed from the EEG data. For this, we used the Gratton and Coles ocular correction algorithm (Gratton et al., 1983). The respective hEOG and vEOG channels were used as reference for eye artefact detection. The data were segmented into 900 ms long epochs starting 300 ms before note onset. Subsequently, an automatic artifact rejection based on the signal from the electrodes of interest Fz, FCz and Cz was performed. The artifact rejection removed all segments that included voltage steps larger than 50 μV/ms, for which the difference between highest and lowest amplitude was more than 100 μV, in which amplitudes were lower than -100 μV or higher than 100 μV, and for which activity was less than 0.1 μV. On average, 12.1 segments per participant were removed (0 to 146 segments, *SD* = 31.7 segments). This left enough segments per participant and condition for the following analyses (see also Table S2 in the Results section).

The interval between 300 and 200 ms before the event was used for baseline correction (for similar procedures in sequential task paradigms, see Herrojo Ruiz et al., 2009; Maidhof et al., 2013). Single-trial data as well as averages for each Event Type were exported per participant.

To calculate the dependent ERP variable, the ERN amplitude, the signal was first pooled at Fz, FCz, and Cz, as at these sites the ERN is typically maximally pronounced, which was also the case in the present study. As participants were allowed to play in their individual tempo, and the latencies of ERNs in sequential tasks are related to movement onset (Maidhof et al., 2013) and thus indirectly to tempo, we expected large peak latency variations between participants which were visible in single-participant data inspection. To include peaks from all participants, we determined the latencies of the maximum negative peak in the averages, in a time window between 130 ms pre- and 130 ms post-event for each Event Type and participant, and latencies of the preceding maximum positive peak in a time window between 180 ms pre-event and the negative peak (for a similar procedure, see Maier et al., 2012). We subsequently calculated the single-trial negative peak measure as the mean signal in the time window 10 ms before to 10 ms after the negative peak latency determined from the averages and calculated the positive peak measures for each trial relative to the average positive peak accordingly. Single-trial ERN peak-to-peak measures were then calculated as the difference between the negative and positive peak values in each segment. We used peak-to-peak measures, as segments might partly overlap in a sequential task, to account for differences in baseline activity which can indeed be seen in Figure 3.

All trials for which the ERN differed by more than two SDs from the mean values per participant and Event Type were removed, which was the case for, on average, 4.9 % of trials (*SD* = 0.5 %, *Maximum* = 5.7 %) per participant. No participant was excluded based on Cook's outlier detection.

***EEG Data Statistical Analyses***

We defined an LME model with ERN amplitude as dependent variable (see above for the general procedure for defining LME models). Event Type served as independent variable, coded as in the behavioral analyses (see Table 1). Random intercepts per participant were included (adding Event Type as random factor led to singular fit error).

Model comparisons were used to check if any model that included Empathy or Expertise explained significantly more variance than the model without these factors (on Bonferroni-corrected Alpha-levels of .025), with the same procedure used in the behavioral analysis.

***Post-Hoc Analyses on Behavioral and EEG Data***

Two post-hoc analyses were conducted to further investigate other influences on error processing. First, we investigated possible differences between corrected and uncorrected errors (replicating Paas et al., 2021), by defining models with the factor Error Correction (-0.5 = uncorrected, 0.5 = corrected) and the dependent behavioral variables IKI and Volume, and the electrophysiological measure ERN amplitude. For this analysis, we originally included 23 participants: the three participants that did not commit enough large errors could be included, but another participant with less than 10 corrected errors had to be excluded. Trials whose values differed more than 2 SD per condition and participant from the mean for the respective dependent variables in the models were excluded. With Cook's Distance, we determined one outlier for the IKI model (remaining *n* = 22, 17-34 years, *M* = 23.4 years, *SD* = 4.3 years; 14 women, 8 men), one outlier for the volume model (remaining *n* = 22, 17-34 years, *M* = 23.4 years, *SD* = 4.3 years; 14 women, 8 men), and two outliers for the ERN model (remaining *n* = 21, 17-34 years, *M* = 23.4 years, *SD* = 4.5 years, 13 women, 8 men). Again, we tested whether adding either Empathy or Expertise led to significantly more explained variance; results of the model comparisons were interpreted on Bonferroni-corrected Alpha levels of .025.

Secondly, we investigated potential effects of the Difficulty of the respective note, the Event Type Frequency for each participant, and the Insecurity with which participants played on IKIs and on the ERN (as the volume data were used to calculate one of the three potential predictors, we omitted this dependent variable from this analysis step). Additionally, we checked if these potential predictors of behavioral and ERP measures themselves differed by Event Type by calculating separate models with the respective variable as dependent variable and Event Type as independent variable (we allowed for random slopes and intercepts per participant). Then we checked whether either of the three potential predictors, when replacing Event Type as independent variable, led to a better model fit for predicting the behavioral IKI or the ERN. The models determined in the main analyses only containing Event Type as predictor were used for comparison. Again, we corrected the Alpha level with Bonferroni correction, this time to $\alpha$ = .017 due to the three model comparisons for each dependent variable.

## Results

Additional statistical results for all models can be found in the supplementary material.

### Expertise & Empathy

Participants spent 7760.38 h on average playing the piano during their lifetime (range 520 h – 24960 h, SD = 7871.76 h, see Figure S3 for a histogram). The sample had an average EQ score of 39.76 points (range 23 – 56 points, *SD* = 9.48 points, see Figure S4 for a histogram).

## Behavioural Data

On average, correct keypresses occurred in 90.70% of all keypresses, small errors in 3.03%, and large errors in 1.44%. All included participants made at least 10 large errors. For detailed information, see Table S2.

### *IKI*

Neither Empathy ($p$ = .458) nor Expertise ($p$ = .606) explained additional variance when added to the model in addition to Event Type. There was a significant effect of Event Type on IKIs, $F(2,14.66) = 18.18$, $p < .001$. Contrast comparisons revealed a significant difference between small and large errors ($p$ = .010, $b$ = 29.39), but not between correct responses and small errors ($p$ = .608, $b$ = 2.30). After a large error keypress, participants took significantly longer (*M* = 371.03 ms, *SD* = 24.23 ms) to press the next key compared to after a small error keypress (*M* = 355.73 ms, *SD* = 13.40 ms), while the IKI after correct actions was comparable (*M* = 359.62 ms, *SD* = 4.42 ms).

### *Volume*

Again, neither Empathy ($p$ = .091) nor Expertise ($p$ = .238) explained significantly more variance when added to the model. We found a significant effect of Event Type, $F(2,15.04) = 15.94$, $p < .001$. Both correct events ($p < .001$, $b$ = 3.22) and large errors ($p < .001$, $b$ = 2.60) resulted in significantly higher volume levels (*M* = 71.55 velocity, *SD* = 0.47 velocity; and *M* = 69.96 velocity, *SD* = 2.00 velocity, respectively) compared to small errors (*M* = 67.58 velocity, *SD* = 1.59 velocity).

***ERN***

For a display of ERPs in the three Event Type conditions, see Figure 3. Adding either Empathy ($p$ = .393) or Expertise ($p$ = .436) to the model did not result in significantly more explained variance compared to a model containing only Event Type as predictor. There was a significant effect of Event Type, $F$(2,18590.00) = 32.28, $p$ < .001. Contrasts revealed significantly lower amplitudes for correct responses ($M$ = -1.03 µV, $SD$ = 0.30 µV) compared to small errors ($M$ = -1.85 µV, $SD$ = 0.84 µV; $p$ < .001, $b$ = 0.90) and significantly higher amplitudes for large errors ($M$ = -3.33 µV, $SD$ = 1.48 µV) compared to small errors ($p$ < .001, $b$ = -1.39).

**Post-hoc Analyses 1: Comparison of Corrected and Uncorrected Errors**

***IKI***

Including Empathy ($p$ = .236) or Expertise ($p$ = .437) did not lead to significantly more explained variance. There was a significant difference between corrected and uncorrected errors, $F$(1,16.34) = 50.40, $p$ < .001, $b$ = 40.10. After correction, participants pressed the following note significantly faster ($M$ = 302.10 ms, $SD$ = 16.15 ms) than after uncorrected errors ($M$ = 382.72 ms, $SD$ = 15.18 ms).

***Volume***

Neither Empathy ($p$ = .284) nor Expertise ($p$ = .161) explained additional variance. There was a significant effect of Correction, $F$(1,21.34) = 78.03, $p$ < .001, $b$ = 9.98. Corrected errors were played significantly more quietly ($M$ = 61.09 velocity, $SD$ = 2.05 velocity) than uncorrected errors ($M$ = 67.98 velocity, $SD$ = 1.58 velocity).

***ERN***

See Figure 4 for a display of ERPs for corrected and uncorrected (small) errors. Neither Empathy ($p$ = .041) nor Expertise ($p$ = .988) explained significantly more variance when added to the model on a Bonferroni-corrected Alpha-level of .025. There was a significant effect of Correction, $F(1,2727.40) = 6.89$, $p = .009$, $b = 0.86$. Corrected errors ($M$ = -2.88 µV, $SD$ = 1.20 µV) were associated with larger ERN amplitudes than uncorrected errors ($M$ = -1.88 µV, $SD$ = 0.86 µV).

**Post-hoc Analyses 2: Event Type Frequency, Difficulty and Insecurity**

Before examining potential effects of the three variables Event Type Frequency, Difficulty and Insecurity on the behavioral and ERP measure of interest (see below), we checked whether Event Type affected the three variables. For this purpose, we calculated three models with the respective variable as dependent and the factor Event Type as independent variable. For M and SD of all three variables, see Table 2. Event Type Frequency was significantly different between conditions, $F(2,45.00) = 2012.10$, $p < .001$. Correct events were more frequent than small errors ($p < .001$, $b = 85.59$), but large and small errors did not differ significantly ($p$ = .155). Event Type also significantly affected Difficulty, $F(2,18595.00) = 593.72$, $p < .001$. Correct Events were associated with lower difficulty values compared to small errors ($p < .001$, $b$ = -8.53), and large errors were associated with higher difficulty ($p = .001$, $b = 1.73$) compared to small errors. Note that this difference is descriptively small and statistical results are influenced by the high number of data points used. Insecurity values also differed significantly between conditions, $F(2,18597.00) = 45.11$, $p < .001$. In accordance with the results for the absolute volume levels (see above), both correct events and large errors led to less deviation from the mean volume than small errors ($p < .001$, $b = -1.65$ and $p < .001$, $b = -2.60$, respectively).

***IKI.*** The analysis of effects of Event Type Frequency, Difficulty and Insecurity on the IKI revealed that Difficulty predicted IKIs, $F(1,13.89) = 7.61$, $p = .015$, $b = 12.96$ (more difficult notes were followed by longer IKIs), as did Event Type Frequency, $F(1,14.24) = 8.72$, $p = .010$, $b = -7.85$ (lower frequencies led to longer IKIs), but Insecurity did not ($p = .127$). In the comparison between the main analysis model (Event Type as factor) and the models with either Difficulty, Event Type Frequency or Insecurity as factor, Event Type as factor led to a better model fit ($AIC_{\text{EventType}} = 172428$) compared to Difficulty, $\chi^2(4) = 130.08$, $p < .001$ ($AIC_{\text{Difficulty}} = 172550$), Event Type Frequency, $\chi^2(4) = 181.78$, $p < .001$ ($AIC_{\text{EventTypeFrequency}} = 172601$) and Insecurity, $\chi^2(4) = 101.89$, $p < .001$ ($AIC_{\text{Insecurity}} = 172521$).

***ERN.*** Event Type Frequency predicted ERN amplitudes, $F(1,16627.00) = 53.55$, $p < .001$, $b = 1.52$ (larger amplitudes for smaller frequencies), but Difficulty ($p = .356$) and Insecurity ($p = .155$) did not. As for the IKI, the model with Event Type as a factor fitted the data significantly better ($AIC_{\text{EventType}} = 131963$) than a model with Difficulty, $\chi^2(1) = 63.55$, $p < .001$ ($AIC_{\text{Difficulty}} = 132025$), Event Type Frequency, $\chi^2(1) = 11.11$, $p < .001$ ($AIC_{\text{EventTypeFrequency}} = 131972$), or Insecurity, $\chi^2(1) = 62.39$, $p < .001$ ($AIC_{\text{Insecurity}} = 132024$).

## Conclusion for Experiment 1

In Experiment 1, we compared the processing of different error types in a piano-playing paradigm. Our results show that ERN amplitudes as well as behavioral measures vary depending on the type of error. Larger ERN amplitudes were observed for large compared to small errors as well as corrected compared to uncorrected errors, while all errors were accompanied by a larger ERN relative to correct responses. Post-error-slowing occurred only for large errors, while small errors were played in a lower volume than large errors and correct keypresses. The results indicate that the action monitoring system does not only differentiate between right and wrong, but also between different degrees of erroneous actions. Moreover, we did not find evidence that other factors related to expectancy affected the ERN and thus the processing of own responses.

## Experiment 2

Observing errors can be just as important as monitoring one's own errors, for example, when musicians play together or teach others. As established above, the mechanisms of processing vicarious actions appear to be similar, albeit not completely identical, compared to those involved in the processing of own actions. Researchers observed a corresponding ERP component, the oERN (Bates et al., 2005; Miltner et al., 2004; van Schie et al., 2004), and increased activity in the mPFC for observed others' errors (see Koban & Pourtois, 2014).

As outlined for own responses above, also the neural response to observed actions can be modulated by expectancy (see Alexander & Brown, 2011), as has been shown for mPFC activity (Schiffer et al., 2014) and the amplitude of a frontocentral oERN-like ERP component (Albrecht & Bellebaum, 2021a; Kobza & Bellebaum, 2013). Recent studies from our lab even suggest that previously observed valence effects for observed actions on this component can be completely attributed to expectancies (Albrecht & Bellebaum, 2021a, 2021b). Therefore, it is questionable whether the component is related to observed error processing, and we will thus subsequently refer to it as observer mediofrontal negativity (oMN). The strong expectancy effect on the oMN amplitude may suggest a functional dissociation between ERN and oMN, with potentially differing effects of error severity on the two components.

As with active error processing, research in observed error processing has so far focused on a binary classification of response accuracy (e.g. Bates et al., 2005; de Bruijn & Rhein, 2012; Kobza & Bellebaum, 2013). The observational data used in this study were taken from the actively performing participants of Experiment 1. We expected to see higher oMN amplitudes for errors than for correct keypresses, as errors were less frequent and thus more unexpected. But, as there was only a slight difference between small and large error frequency in the videos for Experiment 2 and as we assumed that the oMN was mainly driven by the expectancy of the observed response, we suspected to find no difference in oMN amplitude between the error types and thus a different pattern as for own responses in Experiment 1.

There may also be differences between the monitoring of own and observed actions with respect to potential modulators. The observation – and interpretation – of other persons' actions is a social process, which suggests a link to trait empathy that might not be present in own action monitoring. However, previous findings on such a relationship are mixed: Trait empathy effects on error processing have been observed in some studies for an oMN-like component (Bellebaum et al., 2020; Brazil et al., 2011), but not in others (e.g. Clawson et al., 2014). A possible explanation for these mixed findings could be that empathy has an indirect effect on observed response processing: It may modulate expectancy formation regarding others' actions (see also Lockwood et al., 2015), which in turn influences observed response processing (Albrecht & Bellebaum, 2021a, 2021b). Moreover, depending on the nature of the task, social-cognitive processes may be more or less strongly required to form predictions. Consequently, empathy could be relevant for some tasks, but not others (Albrecht & Bellebaum, 2021b).

A factor potentially less relevant for early processing of observed responses is expertise. For observed error monitoring as coded in the oMN, it is not necessary that the respective movement can be performed by the observers themselves (Desmet et al., 2014), While altered ERP responses have been found for expert pianists compared to other musicians and non-musicians after other pianists' errors (Panasiti et al., 2016), the authors found no differences between groups in the oMN-like component for errors vs. correct responses, but larger amplitudes in the following positivity. Based on the reported findings, we expected oMN amplitudes to be modulated by trait empathy but not by expertise.

Unexpectedly, Paas et al. (2021) found that electrophysiological reactions to other-produced tones differed between subsequently corrected and uncorrected errors, although listeners should not know beforehand whether errors would be corrected or not. To further investigate this accidental finding, we compared oMN amplitudes after (visually) observed errors that were either subsequently corrected or not.

As outlined above, oMN amplitude was hypothesized to be primarily driven by expectancies concerning the observed response, which should lead to a different pattern with respect to error severity processing as for own responses. We aimed to support this expected finding and its interpretation with further exploratory analyses. As in Experiment 1 for active responding, we calculated measures that might influence participants' expectancy, namely the event type frequency for each type of observed action and the difficulty of each observed note. We suspected that at least event type frequency, as an established modulator of expectancies (Schiffer et al., 2014; Wang et al., 2015), and maybe also difficulty (see Albrecht & Bellebaum, 2021b) would predict oMN amplitudes as well as, or even better than, event type. We also explored whether the perceived expertise in the observed player affected the monitoring of observed responses. Finally, to directly compare the processing of own and observed actions, we also conducted an analysis including the ERPs from experiments 1 and 2 with factors agency and event type. In this exploratory analysis amplitude differences between the components ERN and oMN were eliminated via z-standardization. As we hypothesized to find differences between small and large errors in the ERN, but not in the oMN, we expected to find a significant interaction between agency and event type.

## Method

### Participants

As in Experiment 1, experienced pianists were recruited via print-material, social media and mouth-to-mouth advertising. Again, a minimum experience of 1500 h was suggested, but lower values were allowed if participants were able to play the respective material by heart (see below). We recorded data from 29 observer participants. Of these, 3 had to be excluded due to technical problems, three because of low performance in the pre- and post-performance test or during the experiment (see below). The remaining 23 participants consisted of 15 cis-gender men and 8 cis-gender women between 18 and 44 years ($M$ = 24.5 years, $SD$ = 6.4 years). One participant was left-handed, 22 right-handed. All participants reported no previous psychological or neurological illnesses, no intake of medication that could affect the nervous system, and had normal or corrected-to-normal vision. Participation was voluntary and participants received compensation of 40€ or course-credit. The study was in accordance with the declaration of Helsinki and approved by the ethics committee of the Faculty of Mathematics and Natural Sciences at Heinrich-Heine-University, Düsseldorf.

### Material

Participants watched videos that were recorded during data acquisition of Experiment 1. In contrast to Experiment 1, participants were required to know the piece by heart to facilitate observation. To limit the time effort and ensure that participants reached a high performance level, we used only one of the six short pieces per participant that were used in Experiment 1. To obtain a large number of trials per condition, we calculated the number of isolated events for each event type and piece. Large errors were the most infrequent event type, so we chose the piece in which the most isolated large errors were made on average. Consequently, we chose 10 videos in which this piece was played (each from a different participant of Experiment 1) that included as many isolated pitch errors as possible and as few other error types as possible (postslips, missed notes, black key notes, pitch errors that deviated more than two white keys from the correct key). In total, participants watched the same piece being played 60 times. Due to a technical error, one of the 10 chosen videos was watched 12 times, 8 videos were watched 6 times each and one video was not watched at all. Originally, the intention was to play each of the 10 videos 6 times. As the order of the videos was randomized, however, and the focus of the study was on the processing of the single notes, this technical error does probably not affect the results of the study. Participants saw 6600 isolated correct notes being played, 290 isolated small errors and 210 isolated large errors (see Table S11 in the supplementary materials).

The expertise of the pianists that played the piece in the video ranged from 936 to 22620 hours ($M$ = 6423.1 h, $SD$ = 6078.8 h). Videos had a resolution of 1280*720 px and a framerate of 60. The videos always started 1 s (or 60 frames) before the first keypress and ended 1 s (or 60 frames) after the last. They were trimmed at the upper and lower side so that only the piano and the moving hand were visible. For practicing, participants received the score notation and auditory recording of the piece before the experiment.

**Experimental Task and Setup**

Similar to the procedure for Experiment 1, the material was sent to participants before testing and they were instructed to practice about 15 minutes a day on average in a tempo that felt comfortable for them. In contrast to Experiment 1, they were, however, instructed to learn only one piece, and this by heart. Participants stated an average practice time of 130.0 minutes ($SD$ = 93.1 minutes, range 44 – 420 minutes). Before the experimental observation task was conducted in the lab, participants were asked to perform the piece themselves on a digital piano

(Casio LK-S450) while MIDI signal was recorded on a connected laptop. The piano was set on mute to avoid additional feedback and to make the conditions as similar as possible to Experiment 1.

For the experimental task, participants sat before a 1920 x 1080 px desktop monitor. Participants were instructed to watch the videos carefully and count the errors made in each of them. The experiment consisted of 60 video presentations (9 different videos; durations between 31-70 s, *M* = 47.8 s, *SD* = 12.8 s), which were played in random order. The videos were embedded in sequences that also contained control questions after each video (see below). For a display of a sequence, see Figure 5.

Participants could start the sequences themselves. After a short fixation cross (500 ms) the video was displayed. Participants received only visual input, the videos were played without sound. A marker was sent to the EEG recording software every 5$^{th}$ observed keypress. Following the videos and another 500 ms fixation cross, participants were asked how many mistakes the observed person had made in this segment. They could freely enter a number and proceed with the Enter key. After another 500 ms fixation cross, participants were asked how experienced in piano playing they believed the observed person to be on a scale from 1 to 10. Again, they could enter a number and proceed with the Enter key. Subsequently, the next sequence came on, which could again be started by the participant. Stimulus presentation and recording was controlled with Presentation (version 22.0, Neurobehavioral Systems, Albany, CA, USA). After completing the experiment, participants were again asked to play the piece on the muted digital piano while MIDI was recorded.

**Assessment of Empathy and Expertise.** We acquired the measures Empathy and Piano Playing Expertise in Experiment 2 in the same way as in Experiment 1.

**Procedure**

Participants received the material to practice the piano pieces used in the experiment via e-mail two weeks before the actual study in the lab. For testing in the laboratory, participants first gave written informed consent to take part in the study. After this, they played the studied piece by heart. Participants subsequently filled out the demographic questionnaire, including Expertise and Empathy measurements, after which EEG electrodes were attached. Participants then completed the actual experiment which lasted around 60 minutes. Finally, the electrodes were removed and participants played the piece again. Participants received either course credit or 40 € as compensation.

**EEG Recording**

EEG measures were recorded in the same way as in Experiment 1, and markers were sent and reconstructed in the same way.

**Data Analyses**

***Behavioral Data of the Pre- and Post-Tests***

All following steps were performed in MATLAB, version R2017b (Mathworks, Natick, Massachusetts, USA). As for Experiment 1, we used the dynamic score matcher algorithm created by Large (1993; see also Palmer & van de Sande, 1993; Rankin et al., 2009) to compare the recorded MIDI signal with the correct score notation for the pre- and post-experiment piano performance. We calculated the accuracy as the percentage of correctly played notes for each participant, separately for the pre- and post-experiment piano performance. If participants restarted playing the piece during the recording, all previous notes were excluded from further analysis. All participants that had an accuracy of less than 50% in both tests were excluded. This was the case for two participants in total.

### *Event Types Used for the ERP Analysis*

We used the same, previously determined relevant notes and events from Experiment 1 for the ERP analysis in Experiment 2. For this purpose, the notes and event types were extracted from the logfiles corresponding to the respective videos shown in Experiment 2. Inclusion criteria for notes were identical to the Experiment 1 analyses, with the exception that we also included notes with less than 30% note accuracy in Experiment 2: high error rates might indicate systematic errors for the players themselves, but do not indicate potential systematic errors of the observer participants. A computer error during testing caused some videos to end too early for seven participants. In only one of them this led to a significant decrease in analyzable segments, and this participant was thus excluded from the analysis.

### *Behavioral Data Assessed During the Experiment and Data Extracted from the Participants of Experiment 1*

We assessed the measure Number of Perceived Errors (as stated by the participants after each sequence) and then calculated the measure Recognized Error Margin as the absolute difference between the Number of Perceived Errors and the actual error number (as calculated from the logfiles of Experiment 1; all error types were included in this measure). One participant of Experiment 2 who scored more than 1.645 *SD* higher (equivalent to a percent rank < 5) than the other participants in the Recognized Error Margin was excluded. Subsequently, the Perceived Expertise of the observed player (as stated by participants after each sequence, see above) and Objective Expertise of the observed player (Expertise measurement calculated for each player from Experiment 1) were determined. Finally, as measures that may affect expectancy of the observed response, we additionally calculated the Difficulty of each note in the piece, i.e. the number of times in percent, that the note was played incorrectly across all 60 sequences, and the

Observed Event Type Frequency, i.e. the number of times (in percent) that each event (correct response, small error, large error) occurred for each observed person. All continuous measures that were considered subsequently as factors in any analysis were scaled to lie between -0.5 and 0. 5 and then mean-centered.

***Behavioral Data Statistical Analysis***

For the analysis of the behavioral data of the pre- and post-test, an LME analysis in R (version 3.5.3) was performed with accuracy as dependent variable and Measurement Time as fixed effect factor. Random intercepts per participants were allowed. For the procedure determining the final model in terms of the random effects structure please refer to the Methods section of Experiment 1.

Then it was checked whether Empathy or Expertise influenced participants' ability to detect errors by defining an LME model that included Recognized Error Margin as dependent variable and Empathy and Expertise as fixed effect factors and that allowed for random intercepts by subject and by observed video.

Additionally, we investigated the relationship between perceived expertise and objective expertise of the observed player. An LME model with perceived expertise as dependent variable and Objective Expertise as fixed effect was defined, which allowed random intercepts and slopes for Objective Expertise by participant and random intercepts by observed video. Then it was examined whether adding either Empathy, Expertise or the Number of Perceived Errors (as stated by participants after each trial) in the respective trial explained significantly more variance by using model comparisons on a Bonferroni-corrected alpha level of .017. If one variable explained more variance, it was added to the model. If two or three variables explained more variance when added to the model one at a time, we systematically checked which model led to the best model fit.

***EEG Data Preprocessing***

First, EEG markers were recoded based on the MIDI data gained in Experiment 1 for each of the observed players by using MATLAB. Subsequently, the markers were imported to Brain Vision Analyzer (Brain Products, Munich, Germany) for EEG data preprocessing which was conducted in the same way as the processing described for Experiment 1. The artefact rejection removed an average of 5.2 segments (range 0-86 segments, *SD* = 17.0 segments).

Segments were also created in accordance with the procedure in Experiment 1, resulting in three Observed Event Types, namely observed correct response, observed small error, and observed large error (only uncorrected errors were included, but we additionally added the event type observed small corrected errors for a post-hoc analyses). Again, single-trial data and averages per Observed Event Type and participant were exported and electrodes Fz, FCz and Cz were pooled.

The component that we call oMN (often referred to as oERN in the literature) occurs later than the ERN in non-sequential tasks, namely 100 to 300 ms after the event (depending on the task, see Bates et al., 2005; Miltner et al., 2004; van Schie et al., 2004). To date, this component has not been investigated in sequential tasks. If, however, earlier ERN peaks in active sequential tasks are related to the earlier onset of the movement relative to key registration compared to non-sequential tasks (Di Gregorio et al., 2022; Maidhof et al., 2013), it is conceivable that the oMN also peaks earlier in sequential tasks, as the observed movement can be detected earlier. Indeed, visual inspection of our data revealed a negativity that seemed to represent action monitoring between -100 and 100 ms around the observed keypress (see Figure 6). In accordance with this, we determined the latencies of the maximum negative peak in a time window between -100 ms pre-event and 100 ms post-event in the average of each participant in each Event Type condition. The preceding positive peak was searched in the time window between -150 ms and the negative peak. Again, single-trial measures were calculated in an area from 10 ms before to 10 ms after the negative and positive peak latency for the respective participant and condition, and peak-to-peak measures for each trial were determined as the difference between the two values. Because peaks were not as pronounced as in the active data and latencies might have varied between trials within participants (due to different playing speeds of the different players in the videos), we additionally investigated the mean amplitude in the time window from -100 to +100 ms with the same model. Results are reported in the Supplementary Material.

We determined outlier trials in which the oMN amplitude differed by more than 2 SDs from the mean values per Observed Event Type and participant. On average, 4.6 % (*SD* = 0.8 %, *Maximum* = 7.3 %) trials per participant were excluded. One subject was removed as outlier for the oMN analysis with Cook's Distance. The remaining sample consisted of 22 participants, 15 men and 7 women, aged between 18 and 44 years (*M* = 24.6 years, *SD* = 6.6 years). After data-cleaning was complete, more than 100 segments remained for each participant and condition (see Table S12).

### *EEG Data Statistical Analyses*

We first defined an LME model with oMN amplitudes as dependent variable and Observed Event Type as independent variable, coded as in the previously described analyses for Experiment 1 (see Table 1). Random intercepts per participant were set (see Methods part for Experiment 1 for the procedure in determining the random effects structure of the model). All continuous measures that were considered as predictors subsequently were scaled to lie between -0.5 and 0.5 and then mean-centered.

As in Experiment 1, Empathy as well as Expertise of the observers were included as additional factors in the models. Model comparisons were used to check if adding one or both explained significantly more variance compared to the model without both measures. If that was the case, the factor was included in the subsequent model analyses. We additionally tested whether adding the Recognized Error Margin, which represented error recognition accuracy and may be affected by attentional factors, as independent variable explained more variance. The model comparisons were interpreted on a Bonferroni-corrected alpha level of .017.

***Post-Hoc Analyses on EEG data***

For our first post-hoc analysis, oMN amplitudes locked to corrected and uncorrected observed (small) errors were compared. For this purpose, an identical model as for the respective analysis in Experiment 1 was used. We then checked if Empathy, Expertise or Recognized Error Margin explained additional variance for the models.

A second additional analysis was conducted to test whether observed error processing (as reflected in the oMN) was better explained by either the Difficulty or the Observed Event Type Frequency of notes than by Observed Event Type. Additionally, we checked if a model involving the Perceived Expertise of the observed person explained significantly more variance than the model with only Observed Event Type as predictor.

***Post-Hoc Analyses comparing active and passive ERP data***

In an exploratory post-hoc analyses, we aimed to compare the effects of the factor Event Type between the ERN and oMN, that is, between the active and observer participants of Experiments 1 and 2. As we were not interested in amplitude differences between the two dependent ERP variables (ERN and oMN), we used z-transformed data of both the active and observer groups. We determined a model with Event Type (correct, small error, large error, coded as in the previous analyses) as within-subject and Agency (active, observer, coded as -0.5 and 0.5, respectively) as between-subject fixed effect factor. Z-standardized single trial ERP amplitudes were set as dependent variables. Random intercepts per participant were allowed. Cook's distance analysis excluded 3 participants, resulting in a total of 41 participants (17 to 44 years, $M$ = 24.0 years, $SD$ = 5.6 years, 18 women, 23 men), 22 in the observer group (18 to 44 years, $M$ = 24.6 years, $SD$ = 6.6 years, 7 women, 15 men) and 19 in the active group (17 to 34 years, $M$ = 23.3, $SD$ = 4.3, 11 women, 8 men). A potential interaction was resolved by determining the Agency effect for the respective conditions correct, small error and large error.

# Results

Additional statistical results for all models can be found in the supplementary material.

## Expertise & Empathy

Participants had a mean Expertise of 4913.09 hours ($SD$ = 4404.90 h, 780 h - 17160 h). The mean EQ score of all participants was 44.00 ($SD$ = 9.93, 17 – 63). Both variables did not differ significantly between the samples of Experiments 1 and 2 (EQ: $t(42) = -1.44$, $p = .156$; Expertise: $t(42) = 1.50$, $p = .141$). For histograms of both variables, see Figure S13 and S14 in the supplementary material.

## Behavioral Data of the Pre- and Post-Test

Participants had an average accuracy of 86.21% in the pre-experimental test ($SD$ = 11.88%), and an average accuracy of 83.80% in the post-experimental test ($SD$ = 16.18%), which did not differ significantly $F(1,22.00) = 1.18$, $p = .289$, $b = -2.42$. This indicates that participants did not learn additionally by watching the 60 repetitions of the piece.

**Behavioral Data Assessed during the Experimental Task**

Regarding the Recognized Error Margin, participants differed on average by 5.60 (*SD* = 1.66) perceived errors from the actual errors in the videos. The difference between perceived and actual errors was not modulated by Empathy ($p = .561$), Expertise ($p = .097$) or an interaction between Empathy and Expertise ($p = .590$). The Recognized Error Margin was calculated as the absolute difference in each sequence between the number of actual and recognized errors. Looking at over- and under-estimation separately, participants underestimated the number of errors in 76.73% of trials on average (*SD* = 20.85%), overestimated the number of errors in 15.39% on average (*SD* = 18.40%), and correctly stated the number of errors in 7.88% on average (*SD* = 4.71%). In sequences in which participants under- or correctly estimated the number of errors, they failed to notice an average of 44.98% of errors (*SD* = 14.15%). In sequences in which participants over- or correctly estimated the number of errors, they noticed on average 46.39% of additional errors (*SD* = 12.73%). The results show that participants were not excellent at recognizing errors. However, this was not dependent on interindividual measures, and the variance between participants was relatively small. To still control for interindividual differences regarding error recognition, we considered the Recognized Error Margin as a variable in our main analysis.

In a model including perceived expertise as dependent and Actual Expertise as independent variable, adding observers' Empathy ($p = .464$) or Expertise ($p = .054$) did not lead to significantly better models, but adding the perceived number of errors did, $\chi^2(2) = 174.30$, $p < .001$, $AIC_{\text{without}} = 4859.10$, $AIC_{\text{with}} = 4688.80$. The actual expertise of the players did not influence the perceived expertise as a main effect ($p = .952$) or in interaction with the Perceived Number of Errors ($p = .071$), but we found a main effect of Perceived Number of Errors, $F(1,1307.63) =$

186.00, $p < .001$, $b = -6.23$. A higher number of perceived errors led to lower perceived expertise.

**EEG data**

ERPs in response to the different Observed Event Types are displayed in Figure 6. For the number of segments included in each condition for the analysis, see table S12. Neither adding Empathy ($p = .681$) or Expertise ($p = .270$) nor the Perceived Number of Errors ($p = .965$) explained any additional variance compared to a model with Observed Event Type as the only predictor. We found a main effect of Observed Event Type, $F(2,11440.00) = 23.13$, $p < .001$. The contrast between observed small errors ($M = -1.04$ µV, $SD = 0.56$ µV) and observed correct keypresses ($M = -0.45$, $SD = 0.19$ µV; $p < .001$, $b = 0.58$) revealed a significant difference, but no difference was found between large ($M = -1.25$ µV, $SD = 0.66$ µV) and small errors ($p = .248$, $b = -0.21$). Calculating the oMN as a mean amplitude between -100 and 100 ms around the event revealed a similar pattern (see supplementary material, section S19).

**Post-Hoc Analysis: Corrected vs. uncorrected**

ERPs for corrected vs. uncorrected errors are displayed in Figure 7. We found that adding Empathy ($p = .187$), Expertise ($p = .105$) or the Perceived Number of Errors ($p = .589$) did not explain any additional variance when added to a model containing Observed Correction as factor. Observed corrected errors elicited an average amplitude of -1.31 µV ($SD = 0.88$ µV), observed uncorrected errors an average amplitude of -1.19 µV ($SD = 0.64$ µV). Observed Correction did not significantly affect the ERPs, $F(1,1576.20) = 0.14$, $p = .705$, $b = 0.09$.

**Post-Hoc Analysis: Frequency and Difficulty**

As all participants viewed the same videos (with small exceptions due to technical problems, see above), we did not calculate statistical differences between Observed Event Types in Observed Event Type Frequency or Difficulty. The mean values are displayed in Table 3: observed correct keypresses were more frequent than both small and large errors, but small errors were only slightly more frequent than large errors. Observed correct keypresses were less difficult than both error types, but error types barely differed in their Difficulty. Both Observed Event Type Frequency, $F(1,1238.90) = 44.34$, $p < .001$, $b = 0.76$ (smaller amplitudes for more frequent events) and Difficulty, $F(1,31985.00) = 14.81$, $p < .001$, $b = 0.76$ (smaller amplitudes for easier notes) had a significant effect on peak-to-peak amplitudes. However, only Observed Event Type Frequency, $\chi^2(1) = 1.94$, $p = .164$, $AIC_{\text{EventType}} = 264808$, $AIC_{\text{EventTypeFrequency}} = 264808$ served as a similarly good predictor as Observed Event Type. Difficulty was a significantly worse predictor, $\chi^2(1) = 31.29$, $p < .001$, $AIC_{\text{Difficulty}} = 264838$. Adding Perceived Expertise to the main analysis model (i.e. the model for the effect of Observed Event Type on oMN amplitude) did not explain additional variance ($p = .436$).

**Exploratory Analysis comparing active and observer ERP data from Experiments 1 and 2.**

Although there was a main effect of Event Type, $F(2,55383.00) = 49.25$, $p < .001$, and a trend effect of Agency, $F(1,172.00) = 3.83$, $p = .052$, our main interest was on the interaction between Agency and Event Type. This interaction was significant, $F(2,55383.00) = 5.91$, $p = .003$. Resolving the interaction, there was no effect of Agency for correct events ($p = .488$) or small errors ($p = .982$), but for large errors $F(1,3570.00) = 9.38$, $p = .002$, $b = 0.13$: the standardized amplitudes of the ERP component related to monitoring were significantly larger for the active compared to the observer group.

## Conclusion for Experiment 2

We studied the processing of different error types in an action observation paradigm in which participants watched videos of others playing the piano. In accordance with previous studies we found larger amplitudes of the oMN for errors vs. correct responses. As in most studies investigating (observed) error processing (see, for example, Miltner et al., 2004; van Schie et al., 2004), valence effects in this study are confounded by low frequencies of errors. Between-condition differences in observed event type frequencies could thus explain the result pattern as well as differences in observed action valence. The focus of the study was, however, on potential effects of error severity, and there was no significant difference between observed small and large, and also not between observed corrected and uncorrected errors. The result pattern thus differed from the one in Experiment 1, where we found error severity effects for own action processing, as well as effects of error correction. The difference between active and observational response monitoring was further supported by an exploratory analysis directly comparing the data obtained in both experiments. This analysis indeed revealed that large errors, but not small errors and correct responses, were processed differently between active and observer participants, indicating that the error type is less influential in observed action processing than in own action processing. As observed small and large errors were similarly frequent, one could assume that these two error types were similarly expected, which may have led to the comparable processing of small and large observed errors.

## General Discussion

Experiment 1 aimed to identify the effect of error severity on behavioral and electrophysiological action monitoring during piano playing. In Experiment 2, we investigated the electrophysiological effect of error severity when observer participants watched videos of pianists playing.

### Error Severity

In line with the hypothesis, Experiment 1 revealed increased ERN amplitudes for large compared to small errors, which, in turn, elicited larger ERN amplitudes than correct responses. Previous research also found a distinction between different error types (Bernstein et al., 1995; Maier et al., 2008; Maier et al., 2012; Maier & Steinhauser, 2016). However, our study is the first to directly test the effect of error severity within one action dimension. Behaviorally, participants showed post-error slowing for large, but not small errors. Overall, we thus found clear effects of error severity. Post-error slowing was reported in some previous studies investigating piano play (Herrojo Ruiz et al., 2009; Paas et al., 2021). The missing post-error slowing after small errors might be attributed to the expertise in our sample: some previous studies showed that expertise reduced or even eliminated post-error slowing (Crump & Logan, 2013; Jentzsch et al., 2014; Loehr et al., 2013; Rachaveti et al., 2020), depending on task demands (Jentzsch et al., 2014). Further, if speed – or keeping a respective tempo – was emphasized, post-error slowing was reduced or not present (Jentzsch & Leuthold, 2006; Loehr et al., 2013). In the present study, participants had to keep the tempo, possibly explaining why no slowing occurred after small errors. Large errors, on the other hand, might have posed more demands with respect to corrective movements and attention, leading to the observed post-error-slowing, possibly due to a reorienting process (see Buzzell et al., 2017; Notebaert et al., 2009; Núñez Castellar et al., 2010).

As found in previous studies, participants played small errors significantly more quietly than correct notes (Herrojo Ruiz et al., 2009; Maidhof et al., 2009; Maidhof et al., 2013; Paas et al., 2021), whereas large errors were played at a similar volume as correct notes. We included only notes that were succeeded by a correct keypress, so for all (‘uncorrected’) errors, hand movements following the error had to be adapted to keep on playing successfully (sequential correction). Together with the finding of post-error slowing, the reduced volume for only small errors suggests that action correction for small errors might start earlier than for large errors (even at keypress), which confirms recent findings on early error movement cancellation effects (Foerster et al., 2022).

For action observation (Experiment 2) there was no difference in processing between small and large errors, but, in accordance with previous studies, a significantly larger oMN amplitude for observed errors compared to correct keypresses was found (see Bates et al., 2005; Bellebaum et al., 2020; de Bruijn & Rhein, 2012; Koban et al., 2010; Miltner et al., 2004; van Schie et al., 2004). Error recognition accuracy, that is, the difference between the number of perceived and actual errors, did not explain additional variance in the model. Thus, even though the null effect does not allow the conclusion that error severity does not affect observed action processing, we assume that the effect is at least reduced in comparison to own errors. This assumption was further supported by an exploratory analysis in which we compared the ERP amplitude pattern between the active (Experiment 1) and observer (Experiment 2) groups. We found that only for large errors, z-standardized amplitude values were significantly larger for the active than for the observer group.

**Empathy and Expertise**

Model comparisons neither justified the inclusion of empathy nor expertise in predicting error processing as reflected in either behavioral or ERP measures, which suggests that these characteristics of the participants did not affect own or observed action monitoring. For empathy in own actions, this is in accordance with previous ERN results (Amiruddin et al., 2017). In contrast, some previous studies found empathy to modulate observed action monitoring (Bellebaum et al., 2020; Brazil et al., 2011; Fukushima & Hiraki, 2009). However, we could show that the effect of empathy on observed response processing was indirect and only occurred because empathy facilitated the formation of expectancies (Albrecht & Bellebaum, 2021a, 2021b). Moreover, this effect emerged in a task with a false-belief component requiring social cognitive abilities. We assume that in the present task without false-belief component participants could form expectancies irrespective of empathy.

Our results do also not support an effect of expertise on own action monitoring that was found in a previous study (Jentzsch et al., 2014). Rachaveti et al. (2020) reported altered post-error slowing after 15 days of practice in a video game. However, both these studies found differences between novices and early stage experts, whereas our sample consisted solely of highly experienced participants (at least 520 hours of experience). There was also no effect of expertise on observed action processing. While a potential expertise effect in observation might be expected in a later time window than the oMN (Panasiti et al., 2016), the missing influence might again be due to an expertise ceiling effect (at least 780 h), as in experiment 1.

**Corrected vs. uncorrected errors**

In our first post-hoc analysis, in accordance with Paas et al. (2021), we found faster post-action reaction times in the active sample after (small) corrected than uncorrected errors. This acceleration probably compensated for the time lost by the correction. As Paas et al., we found that participants played corrected error notes significantly more quietly than uncorrected error notes. Please note that we only included events followed by a correct note. We thus have to differentiate between two processes: immediate note correction (the correct note was repeated after an error; as in the Event Type corrected errors) and sequential correction (the hand movement following the error was adapted so that the subsequent note was played correctly; which happened in all not-corrected errors with following correct notes). Thus, in this case, we would like to also interpret the pattern of lower volumes and smaller post-error reaction times to reflect an earlier onset of corrective actions (either note correction or sequential correction) for corrected (small) errors compared to uncorrected (small) errors. The higher ERN amplitudes for corrected than uncorrected errors also replicate Paas et al.'s findings. In the current study, participants were acting against instructions when correcting errors, as they were told to keep on playing – in the same tempo – if an error occurred. Similarly, in Paas et al.'s joint action study, error correction could negatively impact collaborative play because an additional note was included into the rhythm structure. Error correction seems to be a natural, maybe automatic, process in sequential tasks (Crump & Logan, 2013). In most piano playing scenarios outside of practice and especially in the current task, participants have to actively suppress error correction. In line with accounts that suggest that errors or infrequent events deviate attention away from the task, resulting, inter alia, in post-error slowing (Notebaert et al., 2009; Núñez Castellar et al., 2010), we suspect that participants perceived later corrected errors as especially significant, which led to particularly strong attention deviation and a subsequent failure to suppress correction. This suggests that both subjective importance (corrected vs. uncorrected errors) and physical deviance from the correct actions (large vs. small errors) influences ERN amplitudes. As for the main analyses, the EEG and behavioral effect of correction was neither modulated by empathy nor expertise.

In the observation data, there was no difference between corrected and uncorrected errors in the oMN. In contrast, Paas et al. (2021) found a correction effect for observers – however, an important methodological difference is that participants in their study listened to others playing (i.e., ERPs were locked to observed auditory feedback), while participants in our study watched muted videos of others playing (i.e. ERPs were locked to the observed action). Consequently, Paas et al. found differences at longer latencies in the ERPs (215 to 340 ms). Corrective notes followed very closely to the previous error note, so the component might have coded the accumulation of two consecutively played errors (wrong note + correction) in the case of correction. Our oMN occurred earlier, often before the keypress, and could not have overlapped temporally with the following correction – especially since corrections took on average longer than in Paas et al.'s study (97 ms opposed to less than 30 ms) – so this interpretation seems likely. Our results do not support Paas et al's alternative interpretation that subjective importance (leading to correction) overlaps between active player and observer. As we found reduced volumes for later corrected errors in Experiment 1 and volume changes probably could be noticed when listening, but not when observing a muted video, observer participants could have also reacted to unexpected volume changes in Paas et al's study.

**Potential Expectancy Effects on (Observed) Response Processing and Their Theoretical Implications**

Expectancy and valence are often confounded in action monitoring studies, as errors are less frequent than correct actions (e.g. Falkenstein et al., 1991; Gehring et al., 1993; Miltner et al., 2004; van Schie et al., 2004). Similarly, the three variables that we measured as potentially influencing expectancies in the present study, i.e. Event Type Frequency, Difficulty and Insecurity, differed significantly between Event Types. To rule out effects that can be explained by expectancy, we therefore checked whether one of the variables predicted ERN amplitudes as well as Event Type. According to the PRO model (Alexander & Brown, 2011), all amplitude variance should be accountable to expectancies, and especially Event Type Frequency (Chase et al., 2011; Jessup et al., 2010), but also Difficulty (Albrecht & Bellebaum, 2021b) have been shown to influence action monitoring responses. For the active group of Experiment 1, however, Event Type could not be exchanged for Difficulty, Event Type Frequency, or Insecurity, without a loss of explained variance. This suggests that the effect of Event Type, and thus error severity, cannot be merely attributed to expectancy for active responding, although some influence of expectancies might be conceivable.

Concerning the oMN in observer participants, Observed Event Type Frequency as measure of expectancy served as just as good a predictor as Observed Event Type. As Observed Event Types differed in their frequency, the variables were confounded, which makes the interpretation difficult. However, the frequency differences between small and large errors were negligible compared to the large frequency difference between correct responses and (small and large) errors. We might thus carefully conclude that the observed pattern of oMN amplitudes in observers, differing between errors and correct responses, but not between small and large errors, can be attributed to Event Type Frequencies, which in turn may have influenced expectancies. This interpretation is consistent with findings from our own (Albrecht & Bellebaum, 2021a, 2021b; Bellebaum et al., 2020; Kobza & Bellebaum, 2013) and others' (Schiffer et al., 2014) previous studies in which observed response valence and expectancy were manipulated independently and in which clear expectancy effects on observed response processing were found.

The different findings in action monitoring for action and observation also have theoretical implications. The PRO model states that mPFC activity reflects the (un)expectedness of outcomes and actions rather than their accuracy (Alexander & Brown, 2011; see Gawlowska et al., 2018; Kobza & Bellebaum, 2013; Schiffer et al., 2014; Wessel et al., 2012). However, our data suggest that the monitoring of own actions at least partially reflects the deviation from a (subjective) goal. Based on the finding of different activations depending on the error size, we assume that the action monitoring system sends a general need-to-adapt signal to update not only prediction models, but also action models. The first assumption is in line with the PRO model (Alexander & Brown, 2011), while the second extends it. For predictions, the magnitude of the adapt-signal depends on the prediction error, which has been shown for different event types: For feedback processing, for example, larger ERP amplitudes were found for infrequent compared to frequent feedback, irrespective of feedback valence (Ferdinand et al., 2012). Also, prediction error size modulates trial-by-trial ERP amplitudes in feedback processing (Fischer & Ullsperger, 2013; Ullsperger et al., 2014), suggesting that amplitudes depend on the size of the prediction adaptation. However, the aforementioned studies modulate a signed prediction error, so any effects can be accounted for not only by expectancies, but also by valence, and valence does seem to play an important role in feedback processing (see Proudfit, 2015). We observed a similar effect of prediction error size for the processing of others' actions, when less predicted actions elicited larger oMN amplitudes, irrespective of action valence (Albrecht & Bellebaum, 2021b). We believe that the adapt-signal, or maybe two overlapping adapt-signals, code the magnitude of prediction (as in Albrecht & Bellebaum, 2021b; Ferdinand et al., 2012) and action adaption needed to meet the desired outcome (as in the current study) continuously (rather than dichotomously). This extension of the PRO model could explain the magnitude of adapt-signals for cases where either an action or a prediction model or both have to be updated. Whether action or prediction adaptations are needed highly depends on the task: in observation, if others' movements cannot be influenced (as in our study), an adapt-signal should be sent for prediction errors, but in active performance, especially in a sequential task, the adapt-signal should (also) be highly dependent on the necessity to update action models.

Future studies might test this suggested extension of the PRO model for both own and observed actions by modulating the necessity to adapt movements quickly (sequential vs. non-sequential task) and, especially in observed action, the possibility to adapt actions at all (observation vs. joint-action tasks, see Loehr et al., 2013; Paas et al., 2021). Additionally, the continuous, non-dichotomous, nature of the signal should be tested by introducing multiple valence levels (correct, almost-error, small error, large error etc.) and extending findings on multiple expectancy levels (from highly expected to highly unexpected, possibly by modulating both signed and unsigned prediction errors). To further corroborate the dissociation between expectancy and error severity, participants' expectancy regarding the action should be assessed directly after each trial.

## Conclusion

In conclusion, our results offer first evidence for a continuous error severity coding in the brain during active action processing. Crucially, our results suggest that this effect cannot be (only) attributed to expectancies, suggesting a reliance on a more general need-to-adapt signal in action processing. In contrast, not error severity but possibly expectancy modulated observed action monitoring, which is in line with prediction error coding and updating predictions. The divergent findings between action and observation concerning the effect of error severity might hint at the representation of different continuous need-to-adapt signals in the mPFC, with different signals playing larger roles in action or observation, respectively. This suggested extension of the PRO model (Alexander & Brown, 2011) should be tested empirically by introducing multiple valence and (extending previous research, see Albrecht & Bellebaum, 2021b; Ferdinand et al., 2012) expectancy levels, and manipulating the significance of action adaptation in own and observed action monitoring.

**Acknowledgements**

The authors would like to thank Philipp Rehs for advice and practical assistance with the programming of Experiment 1. The authors would also like to thank Evelina Andronova, Hanna Durdiak, Sonja Lorenz, and Madita Röhlinger for their help with data collection. Further thanks go to Thomas Bartylla for his assistance regarding the experimental stimuli.

**Funding and Conflicts of Interest**

The study was funded by budgetary funds of Heinrich-Heine-University Düsseldorf. The authors report no conflicts of interest.

**Open Practice Statement**

All data, material, and code used in this study as well as the supplementary material are available at osf.io/8e2ja. None of the experiments were preregistered.

## References

Albrecht, C., & Bellebaum, C. (2021a). Disentangling effects of expectancy, accuracy, and empathy on the processing of observed actions. *Psychophysiology*, *58*(9), e13883. https://doi.org/10.1111/psyp.13883

Albrecht, C., & Bellebaum, C. (2021b). Effects of trait empathy and expectation on the processing of observed actions. *Cognitive, Affective, & Behavioral Neuroscience.* Advance online publication. https://doi.org/10.3758/s13415-020-00857-7

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344. https://doi.org/10.1038/nn.2921

Amiruddin, A., Fueggle, S. N., Nguyen, A. T., Gignac, G. E., Clunies-Ross, K. L., & Fox, A. M. (2017). Error monitoring and empathy: Explorations within a neurophysiological context. *Psychophysiology*, *54*(6), 864–873. https://doi.org/10.1111/psyp.12846

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, *34*(2), 163-175. https://doi.org/10.1023/B:JADD.0000022607.19833.00

Bates, A. T., Patel, T. P., & Liddle, P. F. (2005). External Behavior Monitoring Mirrors Internal Behavior Monitoring. *Journal of Psychophysiology*, *19*(4), 281–288. https://doi.org/10.1027/0269-8803.19.4.281

Bellebaum, C., Ghio, M., Wollmer, M., Weismüller, B., & Thoma, P. (2020). The role of trait empathy in the processing of observed actions in a false-belief task. *Social Cognitive and Affective Neuroscience*, *15.* https://doi.org/10.1093/scan/nsaa009

Bernstein, P. S., Scheffers, M. K., & Coles, M. G. H. (1995). "Where did I go wrong?" A psychophysiological analysis of error detection. *Journal of Experimental Psychology. Human Perception and Performance*, *21*(6), 1312–1322. https://doi.org/10.1037//0096-1523.21.6.1312

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (Eds.) (2001). *Conflict monitoring and cognitive control.*: Vol. *108*. American Psychological Association.

Brazil, I. A., Mars, R. B., Bulten, B. H., Buitelaar, J. K., Verkes, R. J., & Bruijn, E. R. A. de (2011). A neurophysiological dissociation between monitoring one's own and others' actions in psychopathy. *Biological Psychiatry*, *69*(7), 693–699. https://doi.org/10.1016/j.biopsych.2010.11.013

Buzzell, G. A., Beatty, P. J., Paquette, N. A., Roberts, D. M., & McDonald, C. G. (2017). Error-Induced Blindness: Error Detection Leads to Impaired Sensory Processing and Lower Accuracy at Short Response-Stimulus Intervals. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(11), 2895–2903. https://doi.org/10.1523/ jneurosci.1202-16.2017

Carter, C. S., Braver T. S., Barch D. M., Botvinick M. M., Noll, D., & Cohen J. D. (1998). Anterior Cingulate Cortex, Error Detection, and the Online Monitoring of Performance.

*Science (New York, N.Y.)*, *280*(5364), 747–749. https://doi.org/10.1126/science.280.5364.747

Chang, A., Chen, C.-C., Li, H.-H., & Li, C.-S. R. (2014). Event-Related Potentials for Post-Error and Post-Conflict Slowing. *PloS One*, *9*(6), e99909. https://doi.org/10.1371/journal.pone.0099909

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience*, *23*(4), 936–946. https://doi.org/10.1162/jocn.2010.21456

Clawson, A., Clayson, P. E., Worsham, W., Johnston, O., South, M., & Larson, M. J. (2014). How about watching others? Observation of error-related feedback by others in autism spectrum disorders. *International Journal of Psychophysiology*, *92*(1), 26-34. https://doi.org/10.1016/j.ijpsycho.2014.01.009

Cracco, E., Desmet, C., & Brass, M. (2016). When your error becomes my error: Anterior insula activation in response to observed errors is modulated by agency. *Social Cognitive and Affective Neuroscience*, *11*(3), 357–366. https://doi.org/10.1093/scan/nsv120

Crump, M. J. C., & Logan, G. D. (2013). Prevention and correction in post-error performance: An ounce of prevention, a pound of cure. *Journal of Experimental Psychology*, *142*(3), 692–709. https://doi.org/10.1037/a0030014

Danielmeier, C., Eichele, T., Forstmann, B. U., Tittgemeyer, M., & Ullsperger, M. (2011). Posterior Medial Frontal Cortex Activity Predicts Post-Error Adaptations in Task-Related Visual and Motor Areas. *Journal of Neuroscience*, *31*(5), 1780–1789. https://doi.org/10.1523/ jneurosci.4299-10.2011

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*(10), 85.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113.

de Bruijn, E. R. A., & Rhein, D. T. von (2012). Is your error my concern? An event-related potential study on own and observed error detection in cooperation and competition. *Frontiers in Neuroscience*, *6*, 8. https://doi.org/10.3389/fnins.2012.00008

de Haen, J. (n.d.). *Deutsche Version der Cambridge Behavior Scale*. http://docs.autismresearchcentre.com/tests/EQ_Deutsch.pdf

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., Cramon, D. Y. von, & Engel, A. K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *25*(50), 11730–11737. https://doi.org/10.1523/ jneurosci.3286-05.2005

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a Neural System for Error Detection and Compensation. *Psychological Science*, *5*(5), 303–305. https://doi.org/10.1111/j.1467-9280.1994.tb00630.x

Desmet, C., Deschrijver, E., & Brass, M. (2014). How social is error observation? The neural mechanisms underlying the observation of human and machine errors. *Social Cognitive and Affective Neuroscience*, *9*(4), 427–435. https://doi.org/10.1093/scan/nst002

Devinsky, O., Morrell, M. J., & Vogt, B. A. (1995). Contributions of anterior cingulate cortex to behaviour. *Brain: A Journal of Neurology*, *118*(1), 279–306. https://doi.org/10.1093/brain/118.1.279

Di Gregorio, F., Maier, M. E., & Steinhauser, M. (2022). Early correlates of error-related brain activity predict subjective timing of error awareness. *Psychophysiology*, *59*(7), e14020. https://doi.org/10.1111/psyp.14020

Eerola, T., & Toiviainen, P. (2004). MIDI toolbox: MATLAB tools for music research.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. Ii. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*(6), 447–455. https://doi.org/10.1016/0013-4694(91)90062-9

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). Erp components on reaction errors and their functional significance: A tutorial. *Biological Psychology*, *51*(2-3), 87–107. https://doi.org/10.1016/s0301-0511(99)00031-9

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

Finney, S. A. (1997). Auditory Feedback and Musical Keyboard Performance. *Music Perception*, *15*(2), 153–174. https://doi.org/10.2307/40285747

Finney, S. A., & Palmer, C. (2003). Auditory feedback and memory for music performance: Sound evidence for an encoding effect. *Memory & Cognition*, *31*(1), 51–64. https://doi.org/10.3758/BF03196082

Fischer, A. G., & Ullsperger, M. (2013). Real and Fictive Outcomes Are Processed Differently but Converge on a Common Adaptive Mechanism. *Neuron*, *79*(6), 1243–1255. https://doi.org/10.1016/j.neuron.2013.07.006

Foerster, A., Steinhauser, M., Schwarz, K. A., Kunde, W., & Pfister, R. (2022). Error cancellation. *Royal Society Open Science*, *9*(3), 210397. https://doi.org/10.1098/rsos.210397

Fu, Z., Wu, D.-A. J., Ross, I., Chung, J. M., Mamelak, A. N., Adolphs, R., & Rutishauser, U. (2019). Single-Neuron Correlates of Error Monitoring and Post-Error Adjustments in Human Medial Frontal Cortex. *Neuron*, *101*(1), 165-177.e5. https://doi.org/10.1016/j.neuron.2018.11.016

Fukushima, H., & Hiraki, K. (2009). Whose loss is it? Human electrophysiological correlates of non-self reward processing. *Social Neuroscience*, *4*, 261–275. https://doi.org/10.1080/17470910802625009

Ganushchak, L. Y., & Schiller, N. O. (2008). Motivation and semantic context affect brain error-monitoring activity: An event-related brain potentials study. *NeuroImage*, *39*(1), 395–405. https://doi.org/10.1016/j.neuroimage.2007.09.001

Gawlowska, M., Domagalik, A., Beldzik, E., Marek, T., & Mojsa-Kaja, J. (2018). Dynamics of error-related activity in deterministic learning - an EEG and fMRI study. *Scientific Reports*, *8*(1), 14617. https://doi.org/10.1038/s41598-018-32995-x

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A Neural System for Error Detection and Compensation. *Psychological Science*, *4*(6), 385–390.

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (Eds.). (2012). *The Oxford Handbook of Event-Related Potential*. *The Error-Related Negativity (ERN/Ne)*. Oxford University Press.

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

Gujing, L., Hui, H., Xin, L., Lirong, Z., Yutong, Y., Guofeng, Y., Jing, L., Shulin, Z., Lei, Y., Cheng, L., & Dezhong, Y. (2019). Increased Insular Connectivity and Enhanced Empathic Ability Associated with Dance/Music Training. *Neural Plasticity*, *2019*, 9693109. https://doi.org/10.1155/2019/9693109

Hajcak, G., McDonald, N., & Simons, R. F. (2003). To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behavior. *Psychophysiology*, *40*(6), 895–903. https://doi.org/10.1111/1469-8986.00107

Hajcak, G., Moser, J. S., Yeung, N., & Simons, R. F. (2005). On the ERN and the significance of errors. *Psychophysiology*, *42*(2), 151–160. https://doi.org/10.1111/j.1469-8986.2005.00270.x

Herrojo Ruiz, M., Jabusch, H.-C., & Altenmüller, E. (2009). Detecting wrong notes in advance: Neuronal correlates of error monitoring in pianists. *Cerebral Cortex (New York, N.Y.: 1991)*, *19*(11), 2625–2639. https://doi.org/10.1093/cercor/bhp021

Hietolahti-Ansten, M., & Kalliopuska, M. (1990). Self-Esteem and Empathy among Children Actively Involved in Music. *Perceptual and Motor Skills*, *71*(3), 1364–1366. https://doi.org/10.2466/pms.1990.71.3f.1364

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

Jentzsch, I., & Leuthold, H. (2006). Short article: Control over speeded actions: A common processing locus for micro- and macro-trade-offs? *Quarterly Journal of Experimental Psychology*, *59*(8), 1329–1337. https://doi.org/10.1080/17470210600674394

Jentzsch, I., Mkrtchian, A., & Kansal, N. (2014). Improved effectiveness of performance monitoring in amateur instrumental musicians. *Neuropsychologia*, *52*, 117–124. https://doi.org/10.1016/j.neuropsychologia.2013.09.025

Jessup, R. K., Busemeyer, J. R., & Brown, J. W. (2010). Error effects in anterior cingulate cortex reverse when error likelihood is high. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(9), 3467–3472. https://doi.org/10.1523/jneurosci.4130-09.2010

Kalfaoğlu, Ç., Stafford, T., & Milne, E. (2018). Frontal theta band oscillations predict error correction and posterror slowing in typing. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(1), 69.

Kalliopuska, M., & Ruókonen, I. (1986). Effects of Music Education on Development of Holistic Empathy. *Perceptual and Motor Skills*, *62*(1), 187–191. https://doi.org/10.2466/pms.1986.62.1.187

Koban, L., & Pourtois, G. (2014). Brain systems underlying the affective and social monitoring of actions: An integrative review. *Neuroscience and Biobehavioral Reviews*, *46 Pt 1*, 71–84. https://doi.org/10.1016/j.neubiorev.2014.02.014

Koban, L., Pourtois, G., Vocat, R., & Vuilleumier, P. (2010). When your errors make me lose or win: Event-related potentials to observed errors of cooperators and competitors. *Social Neuroscience*, *5*, 360–374. https://doi.org/10.1080/17470911003651547

Kobza, S., & Bellebaum, C. (2013). Mediofrontal event-related potentials following observed actions reflect an action prediction error. *The European Journal of Neuroscience*, *37*(9), 1435–1440. https://doi.org/10.1111/ejn.12138

Large, E. W. (1993). Dynamic programming for the analysis of serial behaviors. *Behavior Research Methods, Instruments & Computers*, *25*(2). https://doi.org/10.3758/BF03204504

Larson, M. J., Fair, J. E., Good, D. A., & Baldwin, S. A. (2010). Empathy and error processing. *Psychophysiology*, *47*(3), 415–424. https://doi.org/10.1111/j.1469-8986.2009.00949.x

Lockwood, P. L., Apps, M. A. J., Roiser, J. P., & Viding, E. (2015). Encoding of Vicarious Reward Prediction in Anterior Cingulate Cortex and Relationship with Trait Empathy. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(40), 13720–13727. https://doi.org/10.1523/ jneurosci.1703-15.2015

Loehr, J. D., Kourtis, D., Vesper, C., Sebanz, N., & Knoblich, G. (2013). Monitoring Individual and Joint Action Outcomes in Duet Music Performance. *Journal of Cognitive Neuroscience*, *25*(7), 1049–1061. https://doi.org/10.1162/jocn_a_00388

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

Maidhof, C., Pitkäniemi, A., & Tervaniemi, M. (2013). Predictive error detection in pianists: A combined ERP and motion capture study. *Frontiers in Human Neuroscience*, *7*, 587. https://doi.org/10.3389/fnhum.2013.00587

Maidhof, C., Rieger, M., Prinz, W., & Koelsch, S. (2009). Nobody is perfect: Erp effects prior to performance errors in musicians indicate fast monitoring processes. *PloS One*, *4*(4), e5032. https://doi.org/10.1371/journal.pone.0005032

Maier, M. E., Di Pellegrino, G., & Steinhauser, M. (2012). Enhanced error-related negativity on flanker errors: Error expectancy or error significance? *Psychophysiology*, *49*(7), 899–908. https://doi.org/10.1111/j.1469-8986.2012.01373.x

Maier, M. E., & Steinhauser, M. (2016). Error significance but not error expectancy predicts error-related negativities for different error types. *Behavioural Brain Research*, *297*, 259–267. https://doi.org/10.1016/j.bbr.2015.10.031

Maier, M. E., Steinhauser, M., & Hübner, R. (2008). Is the error-related negativity amplitude related to error detectability? Evidence from effects of different error types. *Journal of Cognitive Neuroscience*, *20*(12), 2263–2273. https://doi.org/10.1162/jocn.2008.20159

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Miltner, W. H. R., Brauer, J., Hecht, H., Trippe, R., & Coles, M. (2004). Parallel brain activity for self-generated and observed errors. In *Errors, Conflicts, and the Brain. Current Opinions on Performance Monitoring* (pp. 124-129). Max Planck Institute for Human Cognitive and Brain Sciences.

Murata, A., & Katayama, J. (2005). An unnecessary response is detected faster than an insufficient response. *Neuroreport*, *16*(14), 1595–1598. https://doi.org/10.1097/01.wnr.0000179080.62529.0b

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, *38*(5), 752–760. https://doi.org/10.1111/1469-8986.3850752

Ninomiya, T., Noritake, A., Ullsperger, M., & Isoda, M. (2018). Performance monitoring in the medial frontal cortex and related neural networks: From monitoring self actions to understanding others' actions. *Neuroscience Research*, *137*, 1–10. https://doi.org/10.1016/j.neures.2018.04.004

Notebaert, W., Houtman, F., van Opstal, F., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: An orienting account. *Cognition*, *111*(2), 275–279. https://doi.org/10.1016/j.cognition.2009.02.002

Núñez Castellar, E., Kühn, S., Fias, W., & Notebaert, W. (2010). Outcome expectancy and not accuracy determines posterror slowing: Erp support. *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 270–278. https://doi.org/10.3758/CABN.10.2.270

Paas, A., Novembre, G., Lappe, C., & Keller, P. E. (2021). Not all errors are alike: Modulation of error-related neural responses in musical joint action. *Social Cognitive and Affective Neuroscience*, *16*(5), 512–524. https://doi.org/10.1093/scan/nsab019

Palmer, C., & van de Sande, C. (1993). Units of knowledge in music performance. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *19*(2), 457–470. https://doi.org/10.1037//0278-7393.19.2.457

Panasiti, M. S., Pavone, E. F., & Aglioti, S. M. (2016). Electrocortical signatures of detecting errors in the actions of others: An EEG study in pianists, non-pianist musicians and musically naïve people. *Neuroscience*, *318*, 104–113. https://doi.org/10.1016/j.neuroscience.2016.01.023

Paulus, C. (2009). *Der Saarbrücker Persönlichkeitsfragebogen SPF (IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index.* http://hdl.handle.net/20.500.11780/3343

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

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

Rabbitt, P. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, *71*(2), 264–272. https://doi.org/10.1037/h0022853

Rabbitt, P. (1969). Psychological refractory delay and response-stimulus interval duration in serial, choice-response tasks. *Acta Psychologica*, *30*, 195–219. https://doi.org/10.1016/0001-6918(69)90051-1

Rabinowitch, T.-C., Cross, I., & Burnard, P. (2012). Long-term musical group interaction has a positive influence on empathy in children. *Psychology of Music*, *41*, 484–498. https://doi.org/10.1177/0305735612440609

Rachaveti, D., Ranganathan, R., & SKM, V. (2020). Practice modifies the response to errors during a novel motor sequence learning task. *BioRxiv*, 2020.10.09.334169. https://doi.org/10.1101/2020.10.09.334169

Rankin, S. K., Large, E. W., & Fink, P. W. (2009). Fractal Tempo Fluctuation and Pulse Prediction. *Music Perception*, *26*(5), 401–413. https://doi.org/10.1525/mp.2009.26.5.401

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science (New York, N.Y.)*, *306*(5695), 443–447. https://doi.org/10.1126/science.1100301

Santesso, D. L., & Segalowitz, S. J. (2009). The error-related negativity is related to risk taking and empathy in young men. *Psychophysiology*, *46*(1), 143–152. https://doi.org/10.1111/j.1469-8986.2008.00714.x

Scangos, K. W., Aronberg, R., & Stuphorn, V. (2013). Performance monitoring by presupplementary and supplementary motor area during an arm movement countermanding task. *Journal of Neurophysiology*, *109*(7), 1928–1939. https://doi.org/10.1152/jn.00688.2012

Schiffer, A.-M., Krause, K. H., & Schubotz, R. I. (2014). Surprisingly correct: Unexpectedness of observed actions activates the medial prefrontal cortex. *Human Brain Mapping*, *35*(4), 1615–1629. https://doi.org/10.1002/hbm.22277

Shane, M. S., Stevens, M., Harenski, C. L., & Kiehl, K. A. (2008). Neural correlates of the processing of another's mistakes: A possible underpinning for social and observational learning. *NeuroImage*, *42*(1), 450–459. https://doi.org/10.1016/j.neuroimage.2007.12.067

Taylor, S. F., Stern, E. R., & Gehring, W. J. (2007). Neural systems for error monitoring: Recent findings and theoretical perspectives. *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *13*(2), 160–172. https://doi.org/10.1177/1073858406298184

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

van Schie, H. T., Mars, R. B., Coles, M. G. H., & Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, *7*(5), 549–554. https://doi.org/10.1038/nn1239

Wang, L., Tang, D., Zhao, Y., Hitchman, G., Wu, S., Tan, J., & Chen, A. (2015). Disentangling the impacts of outcome valence and outcome frequency on the post-error slowing. *Scientific Reports*, *5*, 8708. https://doi.org/10.1038/srep08708

Wessel, J. R., Danielmeier, C., Morton, J. B., & Ullsperger, M. (2012). Surprise and error: Common neuronal architecture for the processing of errors and novelty. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(22), 7528–7537. https://doi.org/10.1523/ jneurosci.6352-11.2012

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931–959. https://doi.org/10.1037/0033-295x.111.4.939

Yoshida, K., Saito, N., Iriki, A., & Isoda, M. (2012). Social error monitoring in macaque frontal cortex. *Nature Neuroscience*, *15*(9), 1307–1312. https://doi.org/10.1038/nn.3180

**Table 1**

*Design Matrix of the factor Event Type*

| | Small Error | Correct | Large Error |
|---|---|---|---|
| **Small Error** | **0.66** | **-0.33** | **-0.33** |
| Correct | -0.33 | 0.66 | -0.33 |
| Large Error | -0.33 | -0.33 | 0.66 |

*Note.* The first line depicts the baseline condition.

**Table 2**

*Means and Standard Deviations for Difficulty, Event Type Frequency and Insecurity by Event Type for Experiment 1.*

| | | |
|---|---|---|
| **Difficulty (in % errors)** | | |
| | *M* | *SD* |
| Correct | 20.02 | 0.22 |
| Small Error | 29.92 | 0.85 |
| Large Error | 30.81 | 1.45 |
| **Event Type Frequency (in % of all events)** | | |
| | *M* | *SD* |
| Correct | 90.33 | 3.34 |
| Small Error | 4.73 | 2.05 |
| Large Error | 2.08 | 1.66 |
| **Insecurity (in velocity, absolute deviation from mean velocity)** | | |
| | *M* | *SD* |
| Correct | 9.68 | 0.14 |
| Small Error | 11.64 | 0.50 |
| Large Error | 8.59 | 0.61 |

**Table 3**

*Means and Standard Deviations for Event Type Frequency and Difficulty by Observed Event Type for Experiment 2.*

| **Difficulty (in % correct)** | | |
|---|---|---|
| | *M* | *SD* |
| Correct | 19.12 | 0.29 |
| Small Error | 27.22 | 1.30 |
| Large Error | 27.70 | 1.32 |
| **Event Type Frequency (in % of all events)** | | |
| | *M* | *SD* |
| Correct | 88.17 | 2.05 |
| Small Error | 4.94 | 1.40 |
| Large Error | 3.31 | 0.79 |

*Note.* M = Mean, SD = Standard Deviation. As all participants saw the same videos (except some minor variation due to technical errors), standard deviations are driven only by differences between videos (and, as for difficulty, differences between notes), but not by differences between participants.

**Figure 1**

*Setup of Experiment 1*

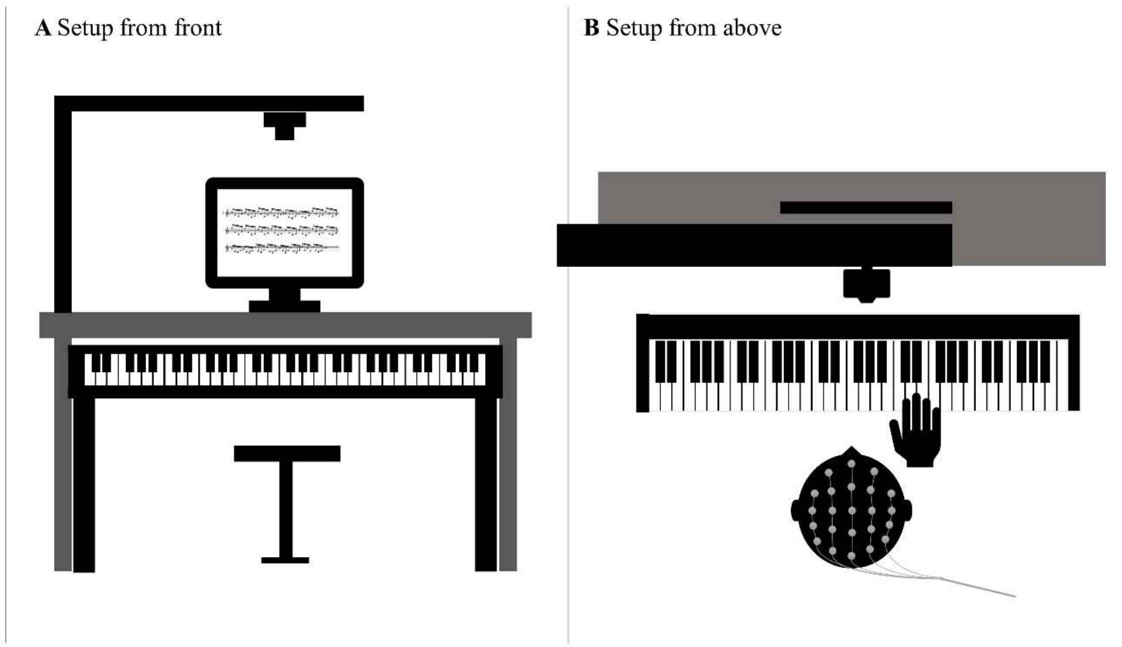


**Figure 2**

*Sequence Structure of Experiment 1*

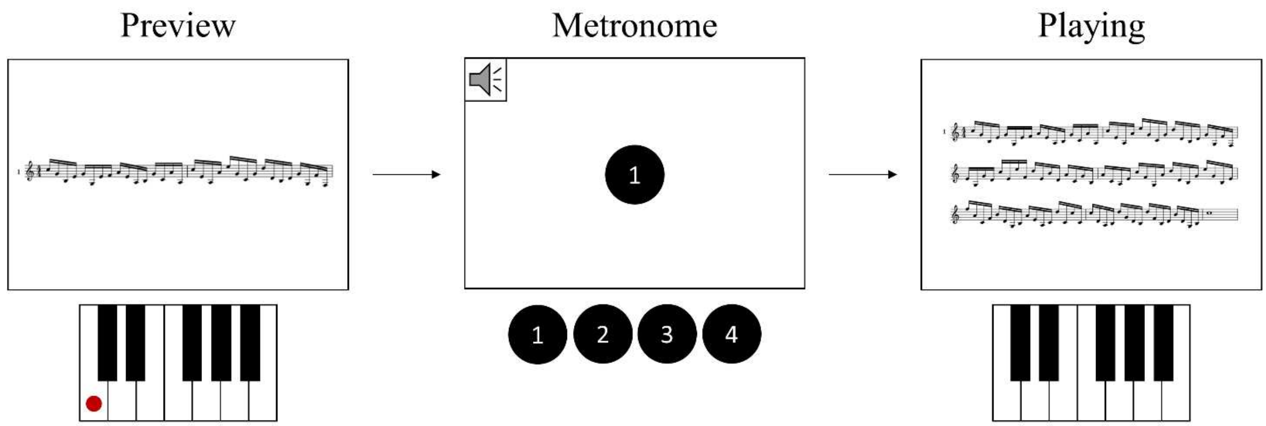
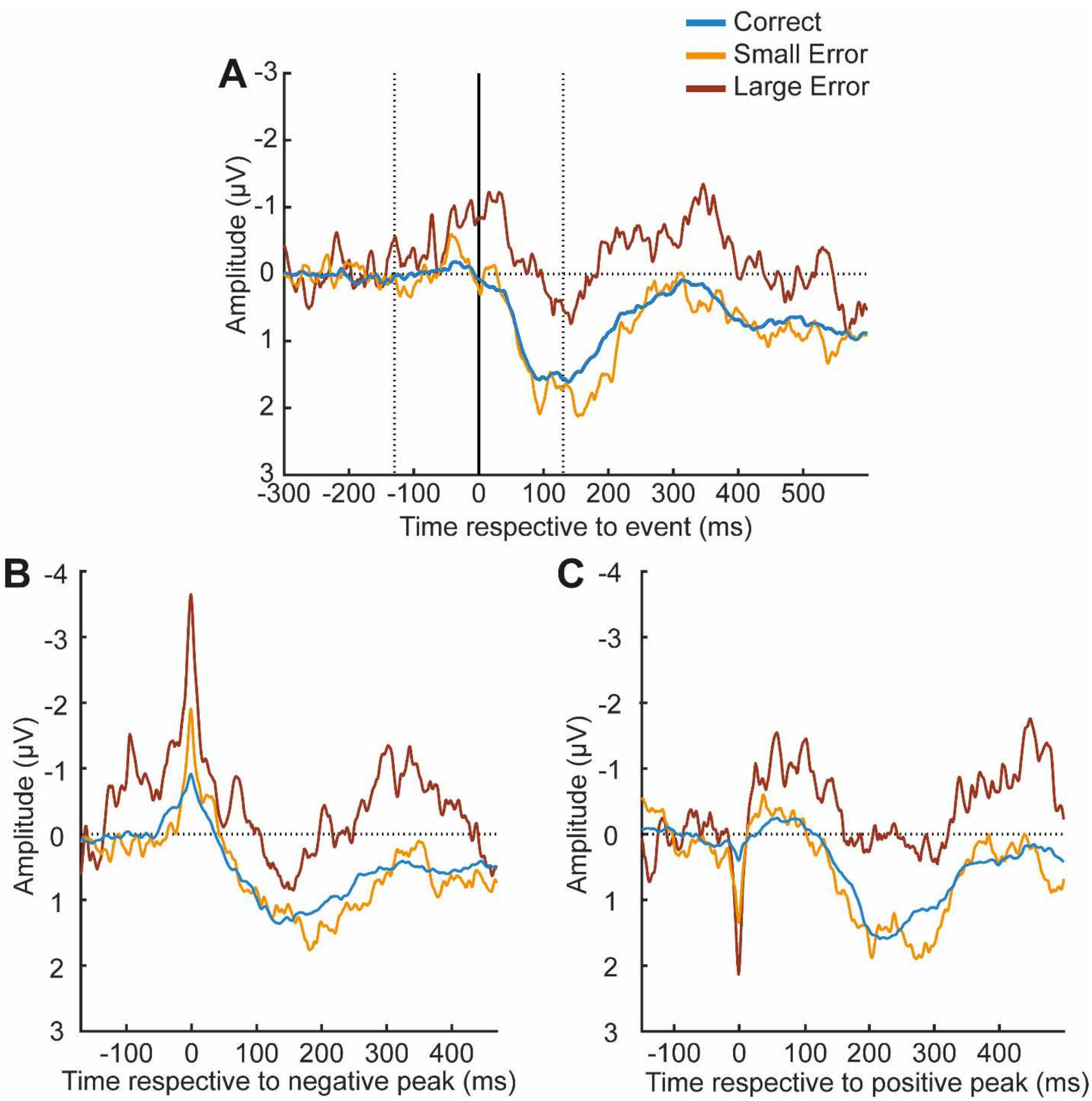


**Figure 3**

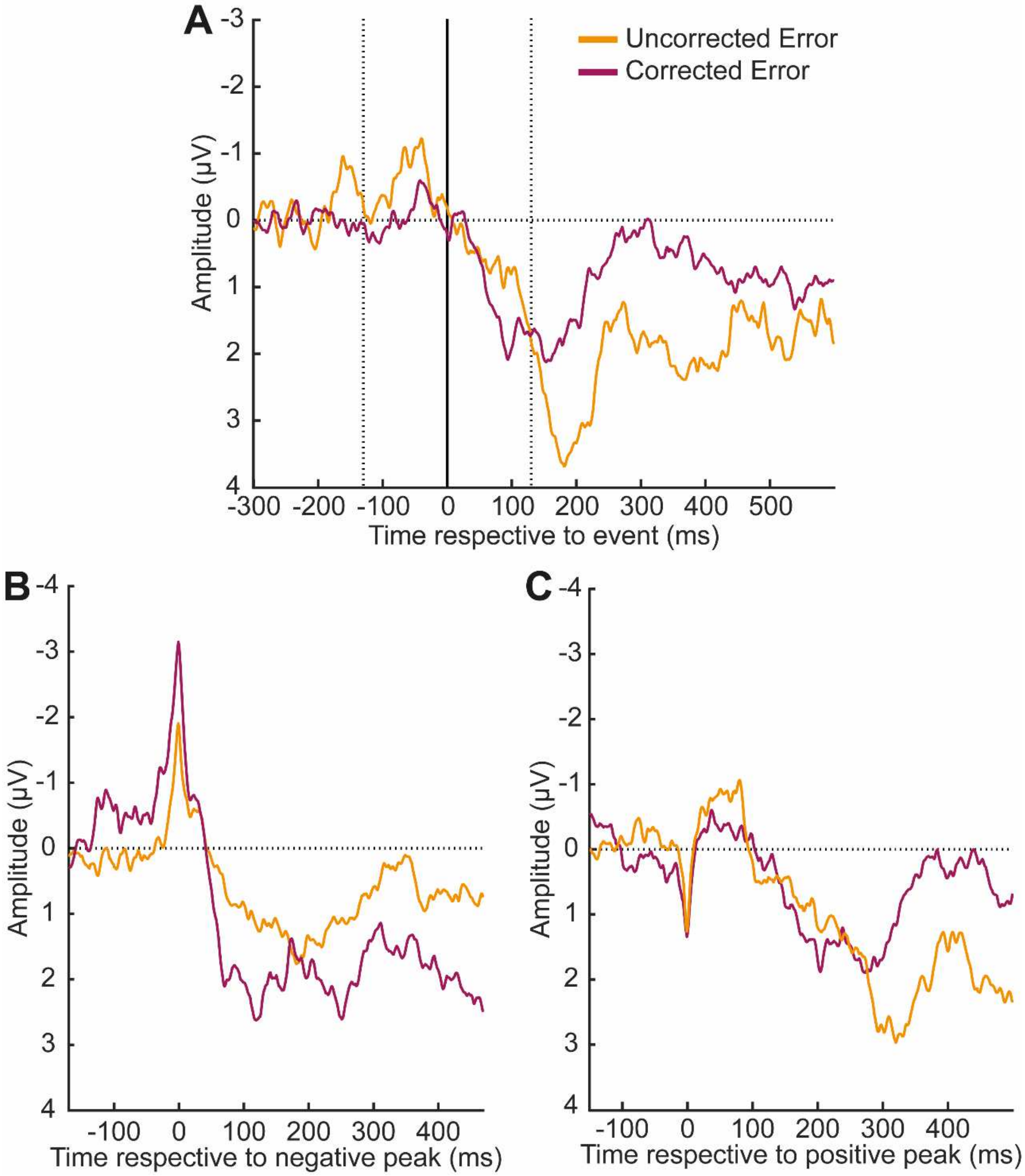
*ERPs as a function of Event Type for Experiment 1*




*Note.* Figure A displays the ERPs respective to the event (correct keypress, small error or large error). Figure B displays the ERPs aligned for the negative peak latency identified for each participant and condition; Figure C displays the ERPs aligned for the respective preceding positive peak latency.

**Figure 4**

*ERPs as a function of Correction for Experiment 1*




*Note.* Figure A displays the ERPs respective to the event (small corrected error or small uncorrected error). Figure B displays the ERPs aligned for the negative peak identified for each participant and condition; Figure C displays the ERPs aligned for the respective preceding positive peak.

**Figure 5**

*Sequence structure in Experiment 2*

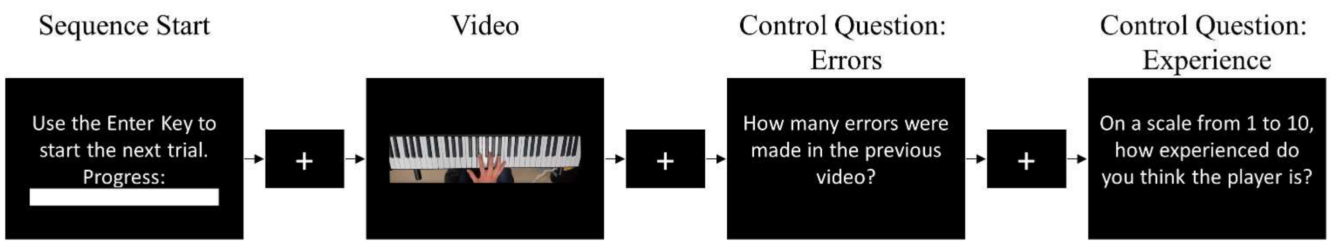
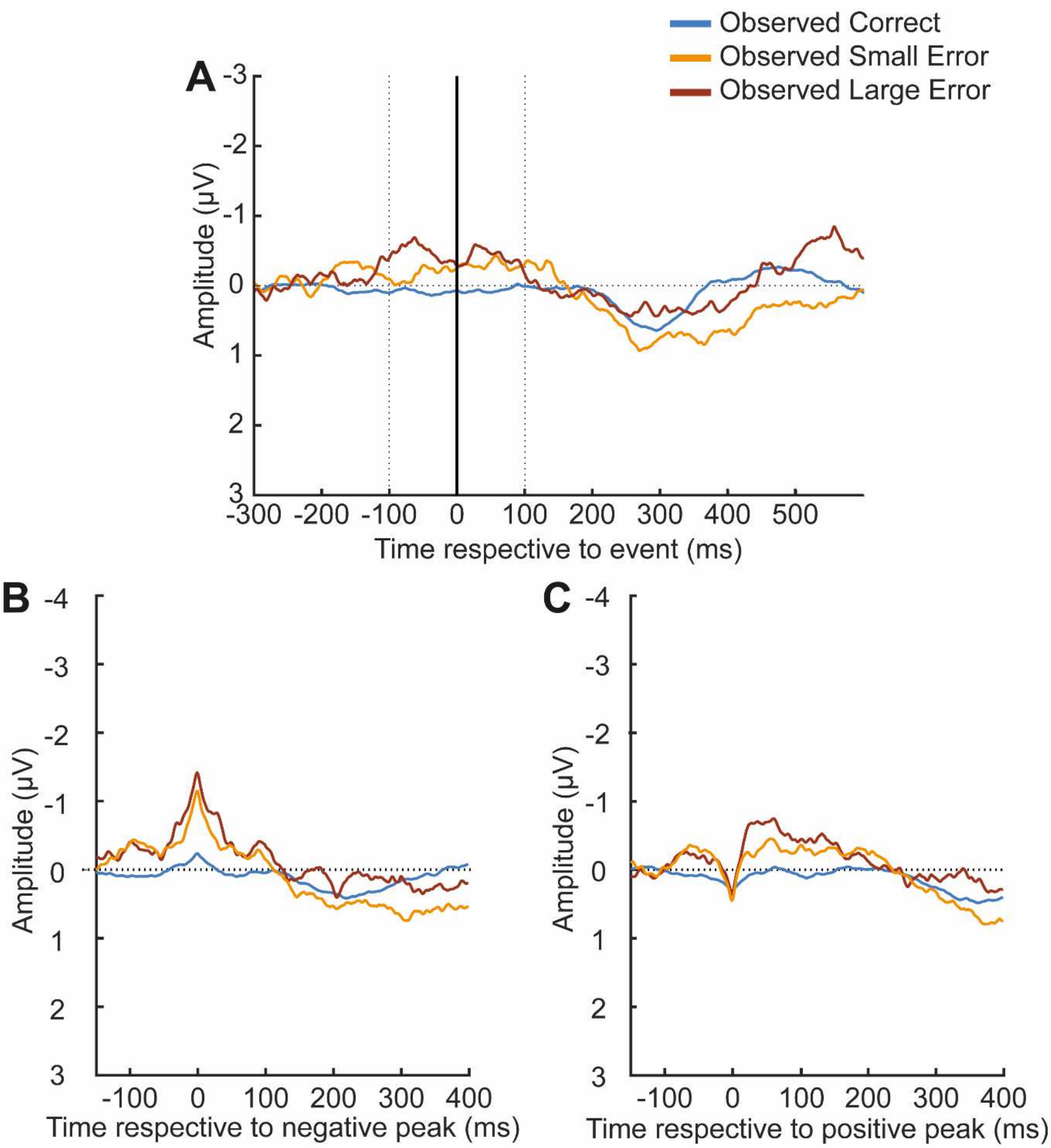


**Figure 6**

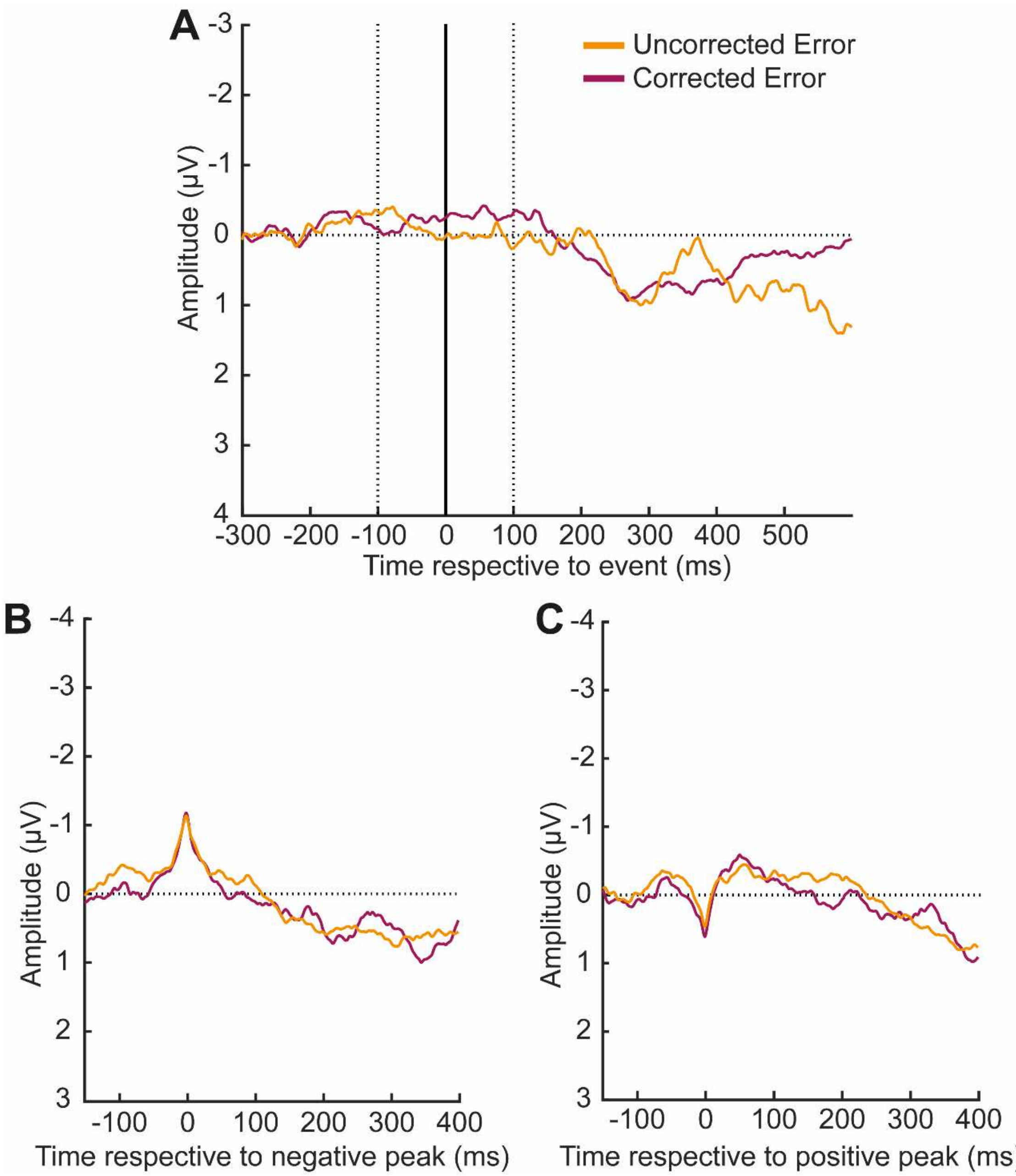
*ERPs of observers as a function of Observed Event Type (Experiment 2).*




*Note.* Figure A displays the ERPs respective to the event (observed correct keypress, observed small error or observed large error). Figure B displays the ERPs aligned for the negative peak identified for each observer participant and condition; Figure C displays the ERPs aligned for the respective preceding positive peak.

**Figure 7**

*ERPs of observer participant as a function of Correction (Experiment 2).*




*Note.* Figure A displays the ERPs respective to the event (small corrected error or small uncorrected error). Figure B displays the ERPs aligned for the negative peak identified for each observer participant and condition; Figure C displays the ERPs aligned for the respective preceding positive peak