Brain and Cognition 99 (2015) 78–86

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c

Effects of invalid feedback on learning and feedback-related brain activity in decision-making

Benjamin Ernst *, Marco Steinhauser

Catholic University of Eichstätt-Ingolstadt, Germany
University of Konstanz, Germany

ARTICLE INFO

Article history:
Received 23 December 2014
Revised 20 July 2015
Accepted 21 July 2015
Available online 8 August 2015

Keywords:
Decision-making
Feedback processing
Event-related potentials
Feedback-related negativity
P300

ABSTRACT

For adaptive decision-making it is important to utilize only relevant, valid and to ignore irrelevant feedback. The present study investigated how feedback processing in decision-making is impaired when relevant feedback is combined with irrelevant and potentially invalid feedback. We analyzed two electrophysiological markers of feedback processing, the feedback-related negativity (FRN) and the P300, in a simple decision-making task, in which participants processed feedback stimuli consisting of relevant and irrelevant feedback provided by the color and meaning of a Stroop stimulus. We found that invalid, irrelevant feedback not only impaired learning, it also altered the amplitude of the P300 to relevant feedback, suggesting an interfering effect of irrelevant feedback on the processing of relevant feedback. In contrast, no such effect on the FRN was obtained. These results indicate that detrimental effects of invalid, irrelevant feedback result from failures of controlled feedback processing.

1. Introduction

Optimal decision-making crucially relies on the ability to improve decisions based on the evaluation of feedback. However, feedback is often ambiguous providing a mixture of valid and invalid information. For instance, a teacher may show an annoyed facial expression before she tells a student that her answer was correct. Even if the student knows that only the oral feedback is relevant and valid, the irrelevant and invalid facial expression might impair learning. The goal of the present study was to investigate whether feedback processing is impaired when relevant and valid feedback is presented together with interfering, irrelevant and potentially invalid feedback. Here, irrelevant feedback is defined as a stimulus that conveys valence information (positive, negative) which, however, is not predictive of the actual future outcome. By considering electrophysiological indices of feedback processing, we aimed to examine whether irrelevant feedback impairs processing of relevant feedback.

In recent years, it has been shown that feedback about the outcome of a simple decision triggers a cascade of event-related potentials (ERPs) that reflect different aspects of learning and feedback processing. The so-called *feedback-related negativity* (FRN) refers to a negative deflection reaching its maximum around 200–300 ms after feedback onset at fronto-central electrode sites, which is more negative for negative feedback than for positive feedback (Miltner, Braun, & Coles, 1997). In their reinforcement learning theory of the error-related negativity (RL-ERN theory), Holroyd and Coles (2002) proposed that the FRN is a correlate of reinforcement learning. According to this account, a negative reward prediction error is generated in the midbrain dopamine system and is conveyed to the anterior cingulate cortex (ACC) where it elicits the FRN and guides learning (Holroyd & Yeung, 2011, 2012). Recently, this account has been modified by assuming that the FRN effect (i.e., the larger negativity following negative feedback) is actually caused by a reward positivity following positive feedback which overlaps with a feedback-locked N200 and which reflects a positive reward prediction error (e.g., Baker & Holroyd, 2011; Foti, Weinberg, Dien, & Hajcak, 2011; Hajihosseini & Holroyd, 2013; Holroyd, Krigolson, & Lee, 2011; Holroyd, Pakzad-Vaezi, & Krigolson, 2008). As an alternative to these variants of the RL-ERN theory, the predicted response-outcome (PRO) model (Alexander & Brown, 2011) proposed that ACC activity

---

* Corresponding author at: Catholic University of Eichstätt-Ingolstadt, Ostenstraße 25, D-85072 Eichstätt, Germany.

E-mail address: Benjamin.Ernst@ku.de (B. Ernst).

http://dx.doi.org/10.1016/j.bandc.2015.07.006

reflected by the FRN represents both positive and negative prediction errors and that differences between positive and negative feedback reflect different degrees of expectedness (Ferdinand, Mecklinger, Kray, & Gehring, 2012). Despite these differences, these accounts share the assumption that the FRN reflects a reward prediction error related to reinforcement learning (for reviews, see San Martín, 2012; Walsh & Anderson, 2012).

A second feedback-related component – the *feedback-locked P300* – is a positivity peaking at posterior electrode sites between 200 and 600 ms after feedback onset. Although most studies found the P300 to be larger for positive feedback (Bellebaum & Daum, 2008; Bellebaum, Polezzi, & Daum, 2010; Ernst & Steinhauser, 2012; Ferdinand & Kray, 2013; Hajcak, Moser, Holroyd, & Simons, 2007; Holroyd, Baker, Kerns, & Müller, 2008; Ulrich & Hewig, 2014; Wu & Zhou, 2009; Zhou, Yu, & Zhou, 2010), others showed a larger P300 for negative feedback (Frank, Woroch, & Curran, 2005; Mathewson, Dywan, Snyder, Tays, & Segalowitz, 2008) or no valence effect at all (Holroyd & Krigolson, 2007; Li, Han, Lei, Holroyd, & Li, 2011; Yeung & Sanfey, 2004; for a review, see San Martín, 2012). The P300 following stimuli in simple decision tasks has been associated with attentional processes or the updating of working memory (Donchin & Coles, 1988; Nieuwenhuis, Aston-Jones, & Cohen, 2005; Polich, 2007). In the context of feedback, the interpretations of the P300 are more specific relating this component to controlled evaluation of action outcomes in working memory (Holroyd & Coles, 2002; Sato et al., 2005; Squires, Hillyard, & Lindsay, 1973; Yeung & Sanfey, 2004; for a review, see San Martín, 2012). Controlled outcome evaluation refers to processes that allow for rapid, flexible behavioral adaptation but rely strongly on attention or processing in working memory (Sailer, Fischmeister, & Bauer, 2010), as proposed in recent models of learning (Collins & Frank, 2012; Frank & Claus, 2006). For instance, a study on reversal learning found that a pronounced feedback-locked P300 predicts correct behavioral adjustment after a contingency reversal, while the FRN reflected adjustments based on prediction errors (Chase, Swainson, Durham, Benham, & Cools, 2011). Despite the ongoing debate about the exact functional significance of these components, the evidence described above suggests that the FRN is more related to reinforcement learning in the ACC, whereas the feedback-locked P300 is associated with controlled feedback evaluation (Chase et al., 2011; Sailer et al., 2010; Walsh & Anderson, 2011). This suggests that feedback is processed by two distinct systems, a perspective that has previously been proposed to account for data in behavioral (Collins & Frank, 2012), fMRI (Daw, Gershman, Seymour, Dayan, & Dolan, 2011), and patient studies (Frank & Claus, 2006).

In the present study, we considered these components to examine how irrelevant and potentially invalid feedback is processed in decision-making. To achieve this, we constructed a simple task in which participants could optimize their decisions (and thus maximize their pay-off) by learning from feedback. The task required participants to decide which one of two characters was associated with a reward. Each character pair was presented a first time in a learning phase and a second time in a test phase. In the learning phase, the decision relied entirely on guessing and feedback had to be evaluated to learn the correct response. Then, in the test phase, correct responding was associated with a reward. In this way, performance in the test phase could be used as an indicator of how efficiently participants utilized feedback in the learning phase. Note that this paradigm differs from reinforcement learning paradigms utilized in previous studies insofar as the stimuli are only presented twice. By this, brain activity following feedback in the learning phase is unaffected by prior learning, but in contrast to gambling tasks this feedback can be used to improve the subsequent performance in the test phase.

Crucially, the feedback stimulus presented in the learning phase was ambiguous[1] and provided two types of feedback. On the one hand, there was a *relevant feedback* that always validly indicated whether the response was correct or not. On the other hand, there was a preceding *irrelevant feedback* that also provided information about the correctness of the response, but this information was valid on half of the trials only. Because this irrelevant feedback contradicted relevant feedback in half of the trials, it was uninformative for learning. Participants knew at any time which feedback was relevant and which was irrelevant, and that only relevant feedback was informative for learning. To ensure that the irrelevant feedback was still processed under these conditions, relevant and irrelevant feedback was realized using Stroop stimuli (Stroop, 1935), that is, colored words whose meaning also referred to a color (e.g., the word BLUE in yellow color). In the present case, the relevant feedback dimension was the word color (e.g., blue for positive feedback, yellow for negative feedback), whereas the irrelevant feedback was the word meaning, which could be valid or invalid depending on whether it referred to the same (e.g., BLUE in blue color) or to the alternative color (e.g., YELLOW in blue color).

A first question was whether the irrelevant feedback has a detrimental effect on learning from relevant feedback, even if participants know that the word is irrelevant. The advantage of using Stroop stimuli is that it is virtually impossible to ignore the word meaning, which is demonstrated by the finding that speeded naming of the color is typically strongly affected by the nature of the word (Stroop effect; for a review, see MacLeod, 1991). However, even if the word is encoded automatically and even if this delays the identification of the color, this does not necessarily imply that it also impairs learning from feedback provided by the color. This is because in contrast to the classical Stroop paradigm there is no response selection under time pressure and no response conflict. Rather, irrelevant feedback could prime either the color category or the feedback valence associated with this color. Accordingly, if irrelevant and relevant feedback are incompatible, this could either impair the identification of the relevant feedback color or activate false valence information. To examine whether this affects learning, we analyzed whether performance in the test phase was impaired if the feedback stimulus in the learning phase contained an invalid word as compared to when it contained a valid word.

If irrelevant feedback has an influence on learning, this should also be reflected in ERPs elicited by the relevant feedback. If irrelevant feedback affects learning more indirectly because it primes color categories and influences the identification of the relevant feedback (which is one component of the Stroop effect; De Houwer, 2003), then both, the FRN and the P300, should be affected by the validity of irrelevant feedback as both rely on accurate stimulus identification. However, if priming occurs on the level of more abstract valence information, it is conceivable that the two components are differentially influenced, with the exact pattern depending on whether this priming influences reinforcement learning or controlled feedback processing or both. On the one hand, if priming occurs on the level of semantic representations of valence (i.e., "right" and "wrong") in working memory, this would affect controlled feedback processing but not necessarily reinforcement learning. In this case, we would expect that irrelevant feedback influences the P300 rather than the FRN. On the other hand, it is possible that controlled processes are better able to protect learning from the influence of irrelevant information. In this case, effects of irrelevant feedback on learning should be

---

[1] Note that, in the present study, only the validity of an irrelevant dimension of the feedback stimulus was manipulated, while the relevant dimension was always valid. This differs from other studies in which feedback validity refers to the correctness of a single feedback stimulus (e.g., Mies, van der Veen, Tulen, Hengeveld, & van der Molen, 2011).