Beyond the Positivity Bias: The Processing and Integration of Self-relevant Feedback is Driven by its Alignment with Pre-Existing Self-Views

Authors:

García-Arch, J.[1,2,3], Friedrich S.[1], Wu, X.[4], Cucurell, D.[1,2,3], Fuentemilla, LL.[1,2,3]

[1]Department of Cognition, Development and Education Psychology, University of Barcelona, Spain
[2]Institute of Neuroscience (UBNeuro), University of Barcelona, Spain
[3]Bellvitge Institute for Biomedical Research, Hospitalet de Llobregat, Spain
[4]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

Abstract

Our self-concept is constantly faced with self-relevant information. Prevailing research suggests that information's valence plays a central role in shaping our self-views. However, the need for stability within the self-concept structure and the inherent alignment of positive feedback with the pre-existing self-views of healthy individuals might mask valence and congruence effects. In this study (N = 30, undergraduates) we orthogonalized feedback valence and self-congruence effects to examine the behavioral and electrophysiological signatures of self-relevant feedback processing and self-concept updating. We found that participants had a preference for integrating self-congruent and dismissing self-incongruent feedback, regardless of its valence. Consistently, EEG results revealed that feedback congruence, but not feedback valence, is swiftly detected during early processing stages. Our findings diverge from the accepted notion that self-concept updating is based on the selective incorporation of positive information. These findings offer novel insights into self-concept dynamics, with implications for the understanding of psychopathological conditions.

Introduction

Individuals hold beliefs about their abilities and attributes that aid in understanding themselves and their environments (Epstein, 1973; Mokady & Reggev, 2022). How these beliefs are formed and updated is a topic that has garnered a lot of attention in recent years. The dominant perspective in this field suggests that when updating self-relevant beliefs, positive and negative information is differently weighted, contributing to the formation of positively biased self-representations (Sharot & Garrett, 2016). While these principles apply to diverse self-relevant beliefs, further considerations are essential to understand self-concept updating. The self-concept is considered a cognitive schema encompassing diverse self-representations, including beliefs about our personality traits (Campbell, 1990; Martinelli et al., 2013). These self-representations are embedded in a highly organized autobiographical knowledge system that protects the self-concept against stability disruptions (Conway, 2005). Consistently, there is evidence that individuals are motivated to seek self-congruent information, regardless of its valence (Swann & Brooks, 2012). This raises questions about the capacity of positive feedback to prompt belief updating independently of its compatibility with pre-existing self-knowledge. Moreover, the inherent positive bias in the self-concept of healthy individuals (Taylor et al., 1988) obscures the distinction between positive and self-congruent information (García-Arch et al., 2023), which might influence the interpretation of findings from previous behavioral and neuroimaging studies. Understanding how individuals form and update self-representations is crucial, since they play a central role in psychological functioning and well-being (Korn et al., 2016; Mokady & Reggev, 2022; Swann et al., 1992). Therefore, unravelling the distinct influences of feedback valence and feedback congruence on self-concept updating requires further inquiry.

Behavioral and neuroimaging studies suggest that desirable and undesirable information is processed and used differently to update self-relevant beliefs, resulting in valence-dependent learning asymmetries (Sharot & Garrett, 2016). Evidence suggest that positive information is readily integrated into our beliefs, while negative information is dismissed (Sharot et al., 2011). The pervasiveness of this phenomenon has led to the assumption that it reflects a fundamental property of learning (Sharot & Garrett, 2016). Recently, these principles have extended to the domain of self-concept updating (Korn et al., 2012, 2014, 2016), consistent with the notion that individuals are motivated to build a positively biased self-view (Hepper et al., 2010). These studies have also shown differential behavioral and neural responses to positive and negative feedback, aligning with a valence-based belief updating bias. Importantly, the propensity towards a valence-dependent updating of self-representations may carry important implications for well-being (Korn et al., 2016; Sharot & Garrett, 2016).

To understand how self-representations might be updated, it is important to consider several important features of the self-concept. Although the self-concept evolves during the lifespan, it also exhibits a pronounced tendency towards stability and coherence (Conway, 2005; Nowak et al., 2000). Self-beliefs, as those related to our personality traits, are well-grounded semantic representations supported by a wide range of autobiographical evidence, which provides certainty and stability to the self-concept (Conway, 2005; Martinelli et al., 2013). We are highly sensitive to information that matches our self-views. Behavioral and neuroimaging studies indicate that we are especially tuned to discern self-related from non-self-related information (Northoff et al., 2006). Information that aligns with our self-perceptions undergoes preferential processing, whereas identity-discrepant information is swiftly identified at the early stages of processing, and subsequently minimized or distorted (Abendroth et al., 2022; Conway, 2005; Nowak et al., 2000). There is also evidence that individuals are motivated to seek self-congruent feedback and protect from self-discrepant evaluations. For example, when facing self-incongruent feedback, individuals experience negative emotional responses, and employ different strategies to mitigate its impact (Swann & Brooks, 2012). Consistently, novel theoretical models suggest that information that matches our self-views might trigger rewarding experiences (Mokady & Reggev, 2022). These findings underscore the pervasive human endeavor to reinforce the certainty and stability of the self-concept. This pursuit aligns with research indicating that a confident and stable self-concept is crucial for daily functioning, bolstering psychological continuity and well-being (Campbell et al., 2003; Jiang et al., 2023; Nowak et al., 2000).

Together, evidence suggests that individuals are motivated to maintain both a positively biased and stable self-concept. However, this dual motivation poses a conceptual challenge in the study of how self-representations are updated. As the self-concept becomes positively biased, positive and self-congruent information converge (García-Arch et al., 2023). This convergence is not trivial, as the distinct behavioral and neural responses elicited by positive and negative feedback might be also explained by variations in its alignment with the existing self-concept. Similarly, different degrees of overlap between feedback valence and self-congruence might produce divergent results across studies and populations. Hence, to unravel the behavioral and neural responses underlying self-relevant belief updating, feedback valence and self-congruence need to be experimentally orthogonalized. Similar concerns have been expressed from different research lines (Mokady & Reggev, 2022; Swann Jr. & Brooks, 2012).

We propose that, in healthy individuals, where a positive bias in the self-concept is already present (Taylor et al., 1988), the drive towards self-concept stabilization would prevail over the need to merely receive positive evaluations. While valence-based belief updating may contribute to building a positively biased self-image, indiscriminate incorporation of positive feedback could undermine self-concept certainty and stability, which are crucial for psychological well-being (Campbell et al., 2003). Note that in healthy individuals, an enhanced focus towards self-concept stability would add certainty to the current self-view at no cost for its overall positivity.

If our hypothesis holds true, feedback that conflicts the existing self-concept should be swiftly identified, which could help avoiding the contamination of self-representations by subjectively inaccurate information (Abendroth et al., 2022). Employing neuroimaging techniques such as the electroencephalography (EEG), with its excellent temporal resolution, can offer critical insights into these processes. The capability of EEG to rapidly distinguish the electrophysiological signatures associated with the processing of feedback valence and congruence may offer novel insights into the dynamics of self-relevant feedback processing.

Here, we required healthy participants to engage in a belief updating task while recording scalp electrophysiological (EEG) activity. Participants evaluated themselves before and after receiving self-relevant social feedback from their peers. We employed a wellknown belief updating paradigm (Elder et al., 2022; Korn et al., 2016) with a recent procedure that allows to control the effect of the initial positive bias in participants self-concept. This procedure allowed us to examine the differential behavioral and electrophysiological signatures associated with the effects of feedback valence and feedback congruence on feedback processing and self-concept updating.

Methods

Participants

Prior to the study, we conducted a power analysis using G*Power (Faul et al., 2007) to determine the required sample size. Following previous literature with similar experimental design (Korn et al., 2012, 2014, 2016), we assumed a partial eta squared of .1 with a conservative correlation between measures of .5. Power analysis revealed that for an acceptable power of .8 20 participants would be required. In the current study, we recruited thirty-five participants (22 females), all of them students from the University of Barcelona. Participants received €10 per hour for participation. Informed consent was obtained from participants following procedures approved by the Ethics Committee of the University of Barcelona. Four participants were excluded because of extensive artifacts in the recorded electroencephalogram (EEG). One participant was excluded due to failing all the attention checks implemented in the experimental task (see details in the next section). The final sample (N = 30, 19 females) consisted of native Spanish speakers; all were right-handed, had normal or corrected-to-normal vision, and had no previous or current neurological or psychiatric disorders. On average, participants were 22.43 years old (SD = 2.17).

Procedure

Participants took part in a two-session experiment separated by 3 days. The first session was online and administered via Qualtrics (www.qualtrics.com). The first session aimed to create a situation in which participants believed they would receive social feedback during the second session. The second session consisted of performing the experimental tasks while EEG was recorded.

First session

This session consisted of an online survey. At the beginning of the survey, participants encountered three embedded audio recordings containing personality descriptions. They were informed that these recordings belonged to anonymous participants contributing to the same experiment within the next 72 hours and had already completed the online survey. Participants' task was to evaluate the speakers' personalities using a provided list of adjectives. To ensure the authenticity of voice samples, recordings were made by independent collaborators who were initially unaware of the aim of the study. Each recording lasted approximately 8 minutes (ranging from 7.45 to 8.29 minutes). After completing their contributions, collaborators were briefed on the study's purpose and provided informed consent for data use. The recordings were presented in random order to the participants. After listening to each personality description, participants evaluated the speaker by choosing applicable adjectives from a predetermined list. Subsequently, they were instructed to record themselves following detailed guidelines and using the presented recordings as examples. These guidelines incorporated 12 randomly chosen items from each of the six HEXACO personality factors (https://hexaco.org/), such as "I feel reasonably satisfied with myself overall" and "I rarely express my opinion in social meetings". Participants were required to speak for at least 30-45 seconds of each statement, expressing their level of agreement and providing contextual examples or anecdotes. Upon completion, they attached their recordings to the online questionnaire.

Next, participants were instructed to evaluate themselves using a list of 150 adjectives (75 positive, see *Stimuli*). The process was designed to control the initial positive bias in participants' self-concept and orthogonalize feedback valence and feedback self-congruence effects. Participants used a drag-and-drop interface to categorize each adjective as "Yes (Me)" or "No (not Me)". They were also instructed to classify adjectives that were unfamiliar to them in an auxiliary box. Adjectives were listed in random order within blocks of positive and negative adjectives, which were also randomized. Participants were instructed to make a minimum of 28 positive and 28 negative decisions, ensuring that negative decisions represented a realistic percentage among the total sample of adjectives (~18%) (García-Arch, et al., 2023). Once this data was obtained, we conducted a non-proportional stratified random sampling on participants positive and negative decisions. That is, the same number of positive and negative decisions were randomly drawn from their respective populations. This strategy allowed us to orthogonalize feedback valence and feedback self-congruence in the session 2.

Finally, participants were requested to complete the Beck Depression Inventory (BDI-II). BDI-II score was used as an exclusion criterion. Following previous research (Garcia-Arch et al., 2022; Kappes & Sharot, 2019), participants that scored >19 in the BDI were excluded from the data analysis. In the current experiment, none of the participants met this criterion.

Second Session

In this session, participants performed a belief-updating task similar to those previously used to study the impact of positive and negative feedback on participant's self-representations (Elder et al., 2022; Korn et al., 2012, 2014, 2016). The task consisted of three blocks: self-evaluation, social evaluative feedback, and re-evaluation phase (Figure 1). In the first block, participants were presented with their own judgments from the first session. Each judgment was displayed on the screen in the format "You think you are [adjective]" or "You think you are not [adjective]", with each adjective (e.g., "Sociable") presented one at a time in random order. Participants were instructed to rate their confidence in each self-assessment using a 0 to 100 slider scale (10 s.), where 0 represented no confidence at all and 100 represented perfect confidence. Participants were instructed to confirm their selection by pressing the space bar within a 10-second interval. The second block introduced social evaluative feedback, purportedly from three other participants who had listened to the participant's voice clip describing their personality. Participants were led to believe that the feedback represented the most frequent judgment among the three evaluators. Each trial began with the question "Do others think you are [adjective]?", that was on the screen for 3 seconds, followed by a fixation cross displayed for a jittered duration of 300, 400, or 500 milliseconds. The evaluators' decision ('Yes' or 'No') was then shown for 1.5 seconds. An inter-trial interval of 1.5 s. including a jittered fixation cross on the screen separated the start of the next trial. Feedback on each adjective was presented three times across three separate blocks, interspersed with rest periods. The feedback was manipulated such that in 25% of cases, participants received positive feedback that matched their self-evaluations (positive + self-congruent), in 25% of cases, the feedback was positive but did not match their self-evaluations (positive + self-incongruent), in another 25% of cases, they received negative feedback that matched their self-evaluations (negative + self-congruent), and in the remaining 25%, the negative feedback did not match their self-evaluations (negative + self-incongruent). We employed categorical feedback to ensure no ambiguities in the perception of its alignment with participants' decisions or its valence. In addition to the main trials, the feedback block also included catch trials to ensure participant engagement and attentiveness. These catch trials followed the same format as the main trials, with the prompt, 'Do others think you are [catch]?', however, in these cases, '[catch]' was replaced with non-adjective words (e.g., “Whistle”). Participants were instructed to identify them by pressing the space bar. After the social evaluative feedback phase, the experiment returned to the initial confidence judgment task (Block 3).

Following the completion of their second session, participants were debriefed. They were informed that the feedback they received was generated pseudo-randomly, and that nobody had actually evaluated their voice clips. They were also informed that the voice recordings they evaluated were made by external collaborators. Additionally, a set of final questions was posed to evaluate any confusion about the stimuli, the task or the setup. No problems were reported.

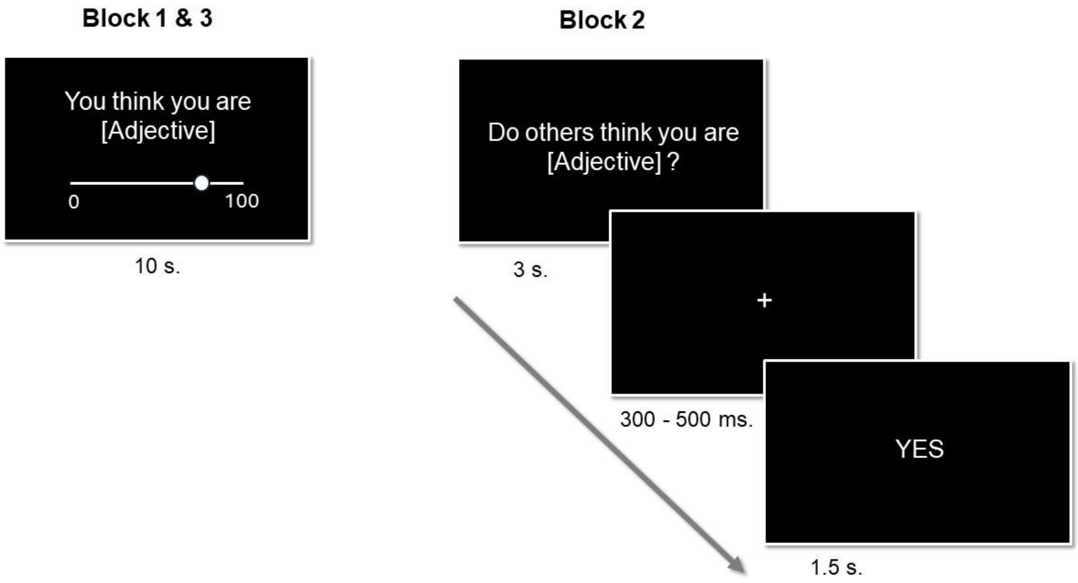
Block 1 & 3
You think you are [Adjective]
0
100
10 s.
Block 2
Do others think you are [Adjective] ?
3 s.
+
300 - 500 ms.
YES
1.5 s.

Figure 1. Overview of the Experimental task. The task is divided into three main blocks. Self-Assessment Rating (Block 1): Participants are presented with statements about their self-judgments from a prior session, formatted as "You think you are [adjective]" or "You think you are not [adjective]." Each adjective is shown individually in a random sequence. Participants rate their confidence in these self-assessments on a 0 to 100 scale, where 0 indicates no confidence and 100 indicates complete confidence. Confirmation of each rating is done via space bar press. Social

Evaluative Feedback (Block 2): Participants receive feedback, purportedly from three peers, on whether others perceive them as described by the adjectives. Feedback is presented in a structured sequence, beginning with a query ("Do others think you are [adjective]?"), followed by a variable-duration fixation cross, the evaluators' decision ('Yes' or 'No'), and another fixation cross before proceeding to the next trial. Feedback is systematically manipulated to include positive and negative evaluations, both congruent and incongruent with the participant's self-assessment. Catch trials with non-adjective prompts are included to monitor engagement and attentiveness. Post-Feedback Reassessment (Block 3): Following the feedback phase, participants revisit the initial confidence rating.

Stimuli

Following previous studies (Elder et al., 2022; García-Arch et al., 2023; Korn et al., 2012, 2014, 2016), we chose personality adjectives to study self-concept updating (i.e., trait words such as 'Sociable', 'Organized', etc.). For the current study, we randomly selected 75 positive (e.g., 'Honest') and 75 negative adjectives (e.g., 'Anxious') from classifications employed in previous studies, which come from widely studied lists of personality descriptors (Anderson, 1968)

Main measures

The target dependent variable for behavioral analysis was *update scores*. These scores represent the change in participants beliefs (i.e., confidence ratings in this study) in the direction suggested by the feedback. That is, post – pre confidence ratings for (positive and negative) congruent feedback and pre – post confidence ratings for (positive and negative) incongruent feedback, representing a measure of 'feedback acceptance' (Korn et al., 2012). All analyses included two binary categorical variables representing the experimental conditions: feedback valence (positive / negative) and feedback self-congruence (self-congruent / self-incongruent). Feedback valence represented whether participants received positive or negative evaluations, while feedback self-congruence was defined by whether those evaluations matched or not participants' decisions. A control measure was included to control for how much space within the scale participants had available for updating (*Update Space*).

EEG recording and preprocessing

EEG was recorded in a faraday cage. Participants were seated in front of the screen at a distance of approximately 57 cm from the center of the screen. The EEG recording was conducted with a 64-channel system at a sampling rate of 250 Hz, using an actiChamp amplifier (Brain Products) and Ag/AgCl electrodes mounted in an electrocap (ANT neuro) located at 60 standard positions (Fp1, Fp2, AF7, AF3, AFz, AF4, AF8, F7, F5, F3, F1, Fz, F2, F4, F6, F8, FT7, FC5, FC3, FC1, FCz, FC2, FC4, FC6, FT8, T7, C5, C3, C1, Cz, C2, C4, C6, T8, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, O2) and the left and right mastoids. One electrode (FT9) was excluded due to technical problems. Eye movements were monitored with an electrode placed at the infraorbital ridge of the right eye. Electrode impedances were kept below 10 kΩ during the recording. FCz served as an online reference. The signal was re-referenced offline to the linked mastoids and bad channels were interpolated (spherical interpolation). A high-pass filter at 0.1 Hz and a low-pass filter at 30 Hz were implemented offline. The continuous EEG data was then epoched into 1s segments. Each epoch spanned a time window from -100 milliseconds (ms) pre-stimulus to 900 ms post-stimulus and a pre-stimulus interval of 100 ms was used as the baseline for absolute baseline correction. Trials exceeding ± 100 μV in EEG and/or EOG channels were automatically rejected offline. Trials containing noise not detected through the amplitude threshold approach were also rejected manually. Preprocessing and statistical analysis of EEG data were conducted in *MATLAB* (Version R2021a) in conjunction with *EEGLAB* (Version 2022.0,Delorme & Makeig, 2004) and *Fieldtrip* (Oostenveld et al., 2011)

EEG data analysis

To investigate the electrophysiological signatures for self-congruent, self-incongruent, positive, and negative feedback, we employed a nonparametric cluster-based permutation test (Maris & Oostenveld, 2007). This data-driven analytical strategy was used to identify clusters of significant points in the spatiotemporal 2D matrix (time and electrodes). This method addresses the multiple-comparison problem by employing a nonparametric statistical testing strategy. The procedure is based on a cluster-level randomization testing to control for the family-wise error rate. Statistics for each time point were calculated, identifying spatiotemporal points with statistical values exceeding a predefined threshold ($p < 0.05$, two-tailed). Next, these points were grouped into clusters based on their adjacency along the x and y axes within the 2D matrix. The observed cluster-level statistics were computed by taking the sum of all values from the contrast statistics within a cluster. Condition labels were then permuted 1000 times (Monte Carlo randomization) to approximate the null hypothesis, and the maximum cluster statistic was chosen to construct a distribution of the cluster-level statistics under the null hypothesis. The significance of the nonparametric statistical test was determined by the proportion of randomized test statistics that exceeded the observed cluster-level statistics. In this analysis, we included the main effects of feedback self-congruence and feedback valence, as well as their one-way 2 x 2 interaction.

Results

Behavioral analysis

Of primary interest, we examined whether participants incorporated more self-congruent than self-incongruent feedback in their self-representations as well as more positive negative feedback. We conducted a repeated measures analysis of variance (rmANOVA) with average update scores as the dependent variable and feedback self-congruence, feedback valence, and their interaction as within-subjects effects. The results of this analysis revealed that participants tended to update significantly more their self-representations in response to self-congruent than in response to self-incongruent feedback ($F(1, 29) = 5.224$, $p = .029$, $\eta_p^2 = .152$). No significant effects were found for feedback valence ($F(1, 29) = 2.435$, $p = .129$, $\eta_p^2 = .077$) and feedback self-congruence x feedback valence interaction ($F(1, 29) = .108$, $p = .743$, $\eta_p^2 = .003$). Next, we aimed to test whether the observed differences in update scores between self-congruent and self-incongruent feedback could be attributed to participants integrating self-congruent feedback (indicated by update scores above zero) and dismissing self-incongruent feedback (reflected by update scores at or below zero), among other possible patterns. Post hoc analysis (one one-sample t-test) revealed confirmed that participants tended to integrate self-congruent feedback into their self-representations (*M* = 3.152, *SE* = .707, 95% *CI*[1.706, 4.598] *t*(29) = 4.458, *p* < .001, *d* = .814) and dismiss self-incongruent feedback (*M* = .334, *SE* = .822, 95% *CI*[-1.346, 2.015] *t*(29) = .407, *p* = .687, *d* = .074) (Figure 2b).

Next, we sought to carry out a more detailed analysis using linear mixed-effects models (LMM). This modelling technique allows to account for individual differences in parameter estimates, include within-subjects covariates (such as update space), compute proper post hoc tests with all the information included in the model, and incorporate additional random effects in the covariance structure of the tested model (Barr et al., 2013; Brown, 2021). We constructed alternative models that varied in their inclusion of fixed effects for feedback self-congruence and feedback valence (each one separate, both main effects, and both with interaction) as well as different combinations of random slopes (see Table S1., *Supplementary Marterials*). All models included partially crossed random effects between adjectives and participants' IDs and update space as a covariate. Model selection was conducted using the Bayesian Information Criteria (BIC), which penalizes model complexity. P-values were determined by Satterthwaite's approximation of degrees of freedom (Kuznetsova et al., 2017). Maximal random effects structures were kept when supported by the data and model convergence (Barr et al., 2013).

Consistent with the rmANOVA results, the winning model (Marginal R2 = .294, Conditional R2 = .409) included feedback self-congruence as a fixed effect, as well as its random slope. The results of this analysis showed that participants tended to incorporate more self-congruent than self-incongruent feedback into their self-representations ($\beta_{\text{Self-congruent}}$ = 18.002, SE = 1.906, 95% CI[14.244, 21.808], t(34.329) = 9.443, p <.001). All models and their associated BICs are reported in Table S1 (*Supplementary Materials*).

EEG results

To investigate the electrophysiological signatures associated with feedback self-congruence and feedback valence, we conducted a cluster-based permutation test on the EEG data recorded during the feedback phase (Figure 1).

The analysis of the EEG data elicited at the feedback cue revealed a significant negative cluster distributed throughout the scalp electrodes between ~300 ms and ~750 ms from cue onset (p = .003, mean t value = -2.825, d = -.515 , peak t value = -5.326, d = -.972), indicating that self-congruent feedback elicited lower ERP amplitudes than self-incongruent feedback (Figure 2, a). No significant clusters were found for the contrasts including feedback valence (Figure 2, c) or feedback self-congruence x feedback valence interaction (all *p* > .124) (Figure S1, *Supplementary Materials*).

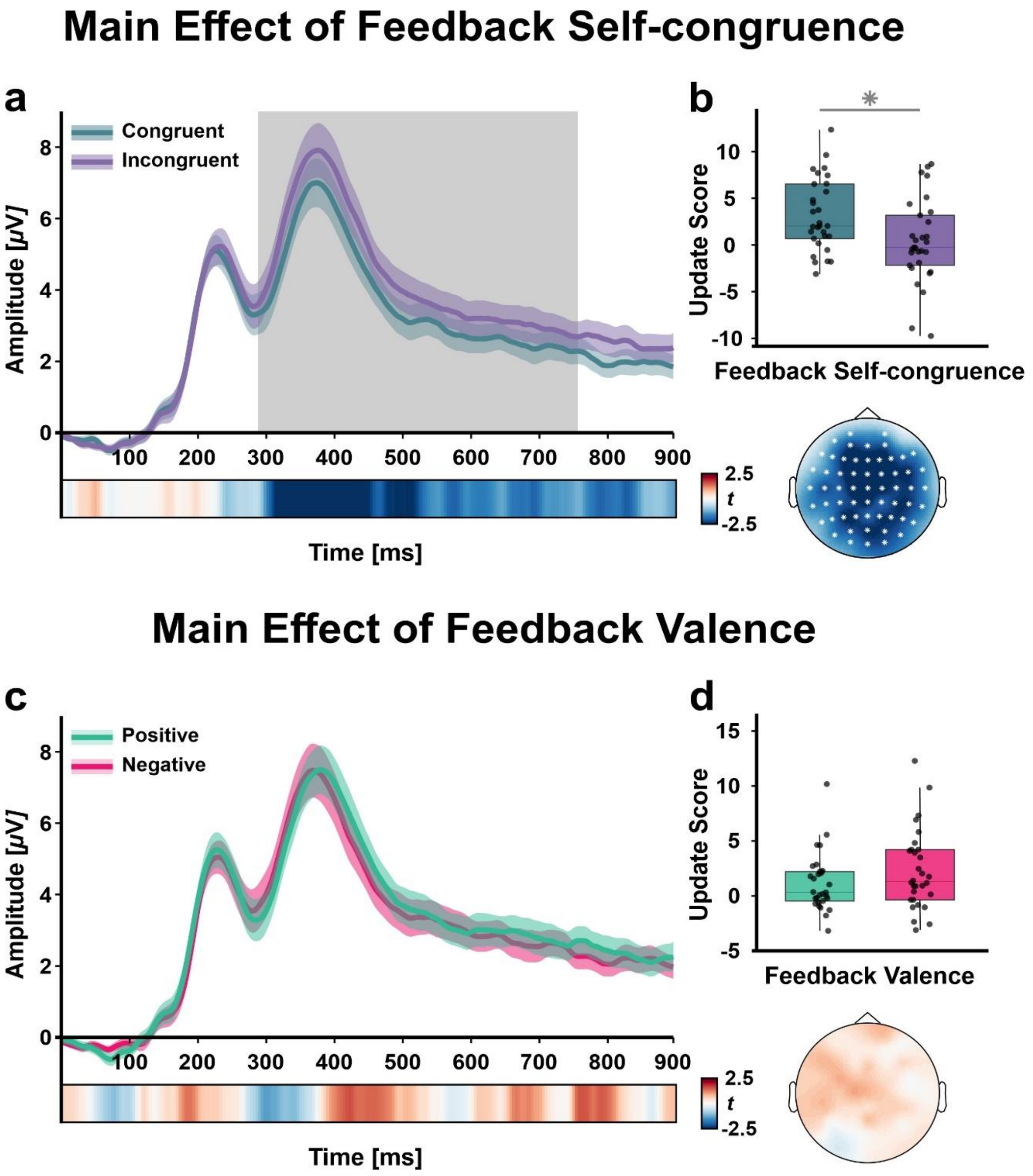

Figure 2: Behavioral and electrophysiological signatures of feedback self-congruence and feedback valence. Panel (a) presents ERP amplitudes in response to congruent (teal blue) and incongruent (purple) feedback over time, with shaded error bands indicating the standard error of the mean. The inset displays the scalp topography of the t-statistic for the main effect of feedback congruence. Panel (b) shows box plots of the main feedback self-congruence on update scores, jittered points represent participants' average. Panel (c) depicts ERP responses to positive (green) and negative (pink) feedback. Panel (d) shows box plots of the main feedback valence on update scores.

Discussion

In this study, we examined the behavioral and neurophysiological responses to social feedback by systematically manipulating feedback valence and self-congruence. Our findings revealed a pronounced asymmetry in the responses to self-congruent and self-incongruent feedback, both at the behavioral and neurophysiological levels. We found that feedback self-congruence was detected at early stages of processing, and that self-congruent information was readily integrated whereas self-incongruent information failed to influence individuals' self-representations. Interestingly, feedback valence did not modulate either behavioral or neurophysiological responses. This finding challenges the widely accepted notion that there is a strong, universal bias towards positive feedback in the updating of self-representations (Korn et al., 2012, 2014). Our experimental orthogonalization of feedback self-congruence and feedback valence provides novel insights into the behavioral and neural signatures of self-relevant feedback processing and self-concept updating.

Our findings revealed a behavioral tendency to selectively assimilate self-congruent and neglect self-incongruent feedback. This is consistent with the notion that self-beliefs are embedded in a rich system of autobiographical information that necessitates mechanisms to stabilize self-representations and protect against conflicting information (Conway, 2005; Nowak et al., 2000). The preferential integration of self-congruent feedback may facilitate the differentiation between self-descriptive and non-self-descriptive attributes, enhancing self-concept clarity (Campbell, 1990). Such clarity in self-concept is crucial for daily functioning, enabling accurate predictions about future behaviors, strategic planning of actions, selection of suitable social partners, and maintenance of psychological well-being (Campbell, 1990; Mokady & Reggev, 2022; Swann & Hill, 1982).

Consistent with our behavioral results, we found that self-congruent and self-incongruent feedback elicited distinct electrophysiological signatures suggesting a rapid discrimination between congruent and incongruent information. Our findings are consistent with ERP literature suggesting that schema-incongruent information triggers rapid electrophysiological responses (Höltje et al., 2019; Richter, 2020).These responses are postulated to reflect a mismatch between incoming information and activated schemas, triggering error signals that result in the updating of mental representations. In contrast, our findings suggested that self-incongruent information did not update participants' self-representations. These differences might be explained by the special nature of the self-concept, which unlike other cognitive schemas is considered to be a highly integrated, emotionally charged structure supported by a lifetime of accumulated evidence (Campbell, 1990; Conway, 2005). These self-concept features promote psychological continuity and might shield self-representations from immediate updates in the face of self-incongruent information (Conway, 2005; Nowak et al., 2000). In line with these notions, recent research suggests that identity-discrepant inputs are detected at early stages of processing and treated as 'false' information (Abendroth et al., 2022) which suggests that the rapid detection of self-incongruent feedback helps protecting self-representations from being disrupted by subjectively inaccurate information.

We did not find significant differences at either the behavioral or electrophysiological level in response to positive and negative feedback. The lack of asymmetry in belief updating, favouring neither positive nor negative feedback, confronts the notion that psychologically healthy adults exhibit a strong tendency to integrate self-relevant information in a valence-dependent manner (Korn et al., 2012). Similarly, we observed no differential electrophysiological responses between positive and negative feedback, diverging from current works that suggest a specialized neural tuning for discerning feedback valence (Korn et al., 2012). These findings may indicate that the convergence of feedback self-congruence with feedback valence —stemming from uncontrolled effects of initial positive biases in individuals' self-concepts— may have masked their effects and led to an overestimation of valence-based effects in previous works.

We suggest that healthy individuals with a positively skewed self-view might have a stronger drive to maintain self-concept stability, which would be compromised if belief updating were driven by unselective integration of positive feedback. Note that reinforcing a positively biased self-concept with confirming evidence would further crystallize self-representations while maintaining its overall positivity. However, we do not dispute the existence of self-related positivity biases. Indeed, the ubiquity of those biases is in itself manifested in the need to control for the initial positive skew in individuals' self-concept to orthogonalize feedback valence and self-congruence. Moreover, although individuals with a positive self-concept seem to prioritize self-concept stabilization, it is possible that this drive towards stability diminishes during pivotal life transitions that require self-concept updates (Conway, 2005). In such instances, a valence-dependent integration of new information might preserve individuals' well-being during adaptive changes.

Our findings may have important implications. The experimental orthogonalization of feedback self-congruence and valence might help reinterpreting findings obtained in previous studies. Moreover, our approach could also improve our understanding of different psychopathological conditions. As a remarkable example, it has been suggested that patients suffering from borderline personality disorder (BPD) display a reduced tendency towards valence-dependent learning asymmetries (Korn et al., 2016). However, this population is also characterized by a more negative self-concept, which can mask congruence and valence effects. Notably, BPD patients are not only characterized by negative self-views, but also by unstable self-concepts (Kaufman & Meddaoui, 2021). Therefore, unravelling congruence and valence effects might help in understanding their neural and behavioral responses to self-relevant information. Finally, the insights extracted from our work could enhance novel approaches based on the modification of maladaptive schemas through schema-incongruent learning in clinical populations (Moscovitch et al., 2023), potentially opening the door to more effective interventions.

Limitations

Following previous works, we focused on the updating of beliefs about personality adjectives. However, the self-concept contains a multiplicity of self-representations such as social roles or group memberships. Future research should extend the current findings to different components of the self-concept.

Acknowledgements

This work was supported by a grant from the Spanish Ministerio de Ciencia, Innovación y Universidades, which is part of Agencia Estatal de Investigación, project: PID2022-140426NB-I00 (funded by MCIN/AEI/10.13039/501100011033/ and FEDER a way to make Europe), to L.F. We thank CERCA Programme/Generalitat de Catalunya for institutional support.

Abendroth, J., Nauroth, P., Richter, T., & Gollwitzer, M. (2022). Non-strategic detection of identitythreatening information: Epistemic validation and identity defense may share a common cognitive basis. *PLoS ONE*, *17*(1 January). https://doi.org/10.1371/journal.pone.0261535

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. In *Journal oj Personality and Social Psychology* (Vol. 9, Issue 3). https://doi.org/https://doi.org/10.1037/h0025907

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/https://doi.org/10.1016/j.jml.2012.11.001

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920960351

Campbell, J. D. (1990). Self-Esteem and Clarity of the Self-Concept. In *Journal of Personality and Social Psychology* (Vol. 59, Issue 3).

Campbell, J. D., Assanand, S., & Di Paula, A. (2003). The Structure of the Self-Concept and Its Relation to Psychological Adjustment. In *Journal of Personality* (Vol. 71, Issue 1, pp. 115–140). https://doi.org/10.1111/1467-6494.t01-1-00002

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, *53*(4), 594–628. https://doi.org/10.1016/j.jml.2005.08.005

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

Elder, J., Davis, T., & Hughes, B. L. (2022). Learning About the Self: Motives for Coherence and Positivity Constrain Learning From Self-Relevant Social Feedback. *Psychological Science*, *33*(4), 629–647. https://doi.org/10.1177/09567976211045934

Epstein, S. (1973). *The Self-Concept Revisited Or a Theory of a Theory*. https://doi.org/https://doi.org/10.1037/h0034679

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

García-Arch, J. , S. A. M. , F. Ll. (2023). Selective Integration of Social Feedback Promotes a Stable and Positively Biased Self-Concept. *Psyarxiv*.

Garcia-Arch, J., Barberia, I., Rodríguez-Ferreiro, J., & Fuentemilla, L. (2022). Authority Brings Responsibility: Feedback from Experts Promotes an Overweighting of Health-Related Pseudoscientific Beliefs. *International Journal of Environmental Research and Public Health*, *19*(22). https://doi.org/10.3390/ijerph192215154

Garrett, N., & Sharot, T. (2017). Optimistic update bias holds firm: Three tests of robustness following Shah et al. *Consciousness and Cognition*, *50*, 12–22. https://doi.org/10.1016/j.concog.2016.10.013

Hepper, E. G., Gramzow, R. H., & Sedikides, C. (2010). Individual Differences in Self-Enhancement and Self-Protection Strategies: An Integrative Analysis. *Journal of Personality*, *78*(2), 781–814. https://doi.org/10.1111/j.1467-6494.2010.00633.x

Höltje, G., Lubahn, B., & Mecklinger, A. (2019). The congruent, the incongruent, and the unexpected: Event-related potentials unveil the processes involved in schematic encoding. *Neuropsychologia*, *131*, 285–293. https://doi.org/10.1016/j.neuropsychologia.2019.05.013

Jiang, T., Wang, T., Poon, K. T., Gaer, W., & Wang, X. (2023). Low Self-Concept Clarity Inhibits Self-Control: The Mediating Effect of Global Self-Continuity. *Personality and Social Psychology Bulletin*, *49*(11), 1587–1600. https://doi.org/10.1177/01461672221109664

Kappes, A., & Sharot, T. (2019). The automatic nature of motivated belief updating. *Behavioural Public Policy*, *3*(1), 87–103. https://doi.org/10.1017/bpp.2017.11

Kaufman, E. A., & Meddaoui, B. (2021). Identity pathology and borderline personality disorder: an empirical overview. *Current Opinion in Psychology*, *37*, 82–88. https://doi.org/https://doi.org/10.1016/j.copsyc.2020.08.015

Korn, C. W., Fan, Y., Zhang, K., Wang, C., Han, S., & Heekeren, H. R. (2014). Cultural influences on social feedback processing of character traits. *Frontiers in Human Neuroscience*, *8*(1 APR). https://doi.org/10.3389/fnhum.2014.00192

Korn, C. W., La Rosée, L., Heekeren, H. R., & Roepke, S. (2016). Social feedback processing in borderline personality disorder. *Psychological Medicine*, *46*(3), 575–587. https://doi.org/10.1017/S003329171500207X

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Martinelli, P., Sperduti, M., & Piolino, P. (2013). Neural substrates of the self-memory system: New insights from a meta-analysis. *Human Brain Mapping*, *34*(7), 1515–1529. https://doi.org/10.1002/hbm.22008

Mokady, A., & Reggev, N. (2022). The Role of Predictions, Their Confirmation, and Reward in Maintaining the Self-Concept. *Frontiers in Human Neuroscience*, *16*. https://doi.org/10.3389/fnhum.2022.824085

Moscovitch, D. A., Moscovitch, M., & Sheldon, S. (2023). Neurocognitive Model of Schema-Congruent and -Incongruent Learning in Clinical Disorders: Application to Social Anxiety and Beyond. *Perspectives on Psychological Science*, *18*(6), 1412–1435. https://doi.org/10.1177/17456916221141351

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain-A meta-analysis of imaging studies on the self. *NeuroImage*, *31*(1), 440–457. https://doi.org/10.1016/j.neuroimage.2005.12.002

Nowak, A., Vallacher, R. R., Tesser, A., & Borkowski, W. (2000). Society of Self: The Emergence of Collective Properties in Self-Structure. In *Psychological Review* (Vol. 107, Issue 1).

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*. https://doi.org/10.1155/2011/156869

Richter, F. R. (2020). *Prediction errors indexed by the P3 track the updating of complex long-term memory schemas*. https://doi.org/10.1101/805887

Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, *20*(1), 25–33. https://doi.org/10.1016/J.TICS.2015.11.002

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479. https://doi.org/10.1038/nn.2949

Swann Jr., W. B., & Brooks, M. (2012). Why threats trigger compensatory reactions: The need for coherence and quest for self-verification. *Social Cognition*, *30*(6), 758–777. https://doi.org/10.1521/soco.2012.30.6.758

Swann, W. B., & Hill, C. A. (1982). When our identities are mistaken: Reaffirming self-conceptions through social interaction. *Journal of Personality and Social Psychology*, *43*(1), 59–66. https://doi.org/10.1037/0022-3514.43.1.59

Swann, W. B., Tafarodi, R. W., Wenzlaff, R. M., & Swann, L. B. (1992). Depression and the Search for Negative Evaluations: More Evidence of the Role of Self-Verification Strivings. In *Journal of Abnormal Psychology* (Vol. 101, Issue 2).

Taylor, S. E., Brown, J. D., Cantor, N., Emery, E., Fiske, S., Green-wald, T., Hammen, C., Lehman, D., McClintock, C., Nisbett, D., Ross, L., & Swann, B. (1988). *Illusion and Well-Being: A Social Psychological Perspective on Mental Health* (Vol. 103, Issue 2).