Selective Integration of Social Feedback Promotes a Stable and Positively Biased Self-Concept

García-Arch, J.[1,2], Sabio Albert, M.[1,2], Fuentemilla Ll.[1,2,3]

[1]Department of Cognition, Development and Education Psychology, Faculty of Psychology, University of Barcelona, Spain.
[2]Institute of Neuroscience (UBNeuro), University of Barcelona, Spain.
[3]Bellvitge Institute for Biomedical Research, Spain

Corresponding author: García-Arch, J., email: j.garcia.arch@ub.edu

Abstract

In daily interactions, we face self-relevant information from our social environment that informs our self-concept. Despite extensive research across various disciplines, there is no consensus on the primary motivations influencing our self-views, with proposals diverging between the pursuit of positive self-images and the need for a stable self-concept. Understanding self-concept dynamics is crucial given its generalized impact in our well-being. However, how we integrate information into our self-representations to promote a positively biased, yet progressively stable self-concept is a question that remains unanswered. In a series of 4 experiments (Experiment 1, n= 33; 2, n= 40; Experiment 3, n= 45; Experiment 4, n= 40), we combined a sentence verification task (Experiments 2-4) with a belief updating task to investigate how participants integrate social feedback depending on its valence and self-congruence. Experiment 1 indicated that the lack of control of an initial positive bias in participants self-concept might have masked valence and congruence effects in recent works. After implementing methodological adjustments (Experiments 2-3) our results suggested that the integration of social feedback was strongly driven by feedback self-congruence and moderately driven by feedback valence. Importantly, both effects showed to be enduring after 24 hours (Experiment 3) and self-specific (Experiment 4). By synthesizing insights from social, personality, and cognitive psychology, this study offers a nuanced understanding of self-concept dynamics during social feedback processing. Our conceptual and methodological advancements have implications for understanding the link between structural and affective components of self-concept and offer a new lens for reinterpreting previous empirical studies.

Introduction

Our self-concept contains information about our personality and life experiences that help us define who we are and what we can expect from ourselves (Conway et al., 2004; Epstein, 1973; Grilli & Verfaellie, 2014; Martinelli et al., 2013; Rathbone et al., 2008). However, the self-concept is not static, and the formation of self-representations is a continuous process (Conway, 2005; Manzi et al., 2010; Markus & Wurf, 1986). This process of construction and revision of self-representations is particularly influenced by social feedback (Crone et al., 2022; Reitz et al., 2014; Rodman et al., 2017). During social interactions, we are often confronted with feedback about our attributes, which serves us to shape the way we see ourselves. Numerous studies suggest that we achieve and maintain a positive self-concept by seeking positive evaluations and selectively integrating them into our self-representations (Alicke & Sedikides, 2009; Hepper et al., 2011; Korn et al., 2012; Taylor et al., 1988). From this perspective, the self-concept is portrayed as a motivated system that selectively incorporates information that maximizes its positivity, even at expense of accuracy (Alicke & Sedikides, 2009; Sedikides & Alicke, 2019). In contrast, research also suggests that our self-representations are embedded in a highly organized self-knowledge system that protects the self-concept from long-term stability violations (Conway, 2005; Conway et al., 2004; Conway & Pleydell-Pearce, 2000; Grilli, 2017; Rathbone et al., 2008). Accordingly, there is also evidence that that we tend to reject self-discrepant feedback (Swann Jr. & Brooks, 2012; Swann et al., 1984; Swann & Hill, 1982) and strive to receive social inputs that are consistent with our self-representations, even if they are negative (Swann Jr. & Buhrmester, 2012; Swann, Tafarodi, et al., 1992). Understanding self-concept stability-malleability dynamics is crucial, given its generalized impact on our cognition, behavior and affect (Beck et al., 1990; Libby & Eibach, 2002; Marsh & Martin, 2011). However, how we incorporate self-relevant information to promote a positively biased but progressively stable self-concept is a question that remains unanswered.

On one side of these opposing views, there is a broad consensus that we are primarily motivated to achieve and maintain a positively biased self-concept. For example, we often believe to possess more favorable traits than others and tend to overestimate our abilities and positive attributes (Dunning et al., 2003; Preuss & Alicke, 2009; Zell et al., 2019). The impact of our self-concept in our well-being is critical and pervasive, and it has been suggested that the motive for achieving and preserving positive self-representations is inherent to psychologically healthy adults (Hepper et al., 2010). Extensive research has also suggested that we are highly skilled in processing information in favor of ourselves (Boseovski, 2010; Taylor et al., 1988). In social interactions, we strive to receive positive feedback and prefer interaction partners who are more likely to provide it (Hepper et al., 2010). We overestimate the likelihood of receiving positive evaluations from others and devote efforts to restore positive self-representations after facing threatening feedback (Rodman et al., 2017). Recently, a growing body of experimental research has suggested that we achieve a positively biased self-concept through the selective incorporation of positive over negative information. Specifically, this research has revealed that when facing new self-relevant social feedback, negative inputs tend to be largely disregarded, whereas positive feedback prompts a shift in self-representations towards a feedback-consistent direction (Elder et al., 2022; Korn et al., 2012, 2014). This feedback-based valence asymmetry in updating the self-concept has provided experimental results with large effect sizes, it has shown to be cross-culturally invariant (Korn et al., 2014) and important for psychological well-being (Korn et al., 2016). These studies suggest that our bias towards positive feedback extends beyond mere preference or active seeking behaviour; it reveals that we selectively utilize positive feedback to shape and enhance our self-representations.

Although the view of the self-concept as a system that tries to maximize its positivity is widely accepted, this conceptualization overlooks important features of self-representations that have been extensively studied in other fields of research in psychology. For example, from a cognitive perspective, self-representations are conceived as an endurable and well-

grounded form of personal semantic knowledge embedded in a highly structured system of autobiographical information (Conway, 2005; Haslam et al., 2011; Klein, 2010; Rathbone & Moulin, 2014; Renoult et al., 2012). There is evidence that our self-representations are formed, supported, and contextualized by a wide range of episodic memories we have encoded from extended, repeated, and self-defining events (Clare J. Rathbone & Conway, 2009; Conway, 2005; Rathbone et al., 2008). This highly organized set of self-relevant information provides stability and coherence and allows us to remember and anticipate trait-congruent experiences and select adaptive behaviors (Conway et al., 2004). From this view, our need to sustain self-concept stability and coherence operates as a powerful constraint that determines which upcoming external inputs would be encoded and remembered, favoring self-congruent information (Conway, 2005; Conway et al., 2004; Conway & Pleydell-Pearce, 2000). Accordingly, social psychology studies have revealed that in social interactions we strive to receive feedback that is congruent with our self-beliefs (Robinson & Smith-Lovin, 1992; Swann & Brooks, 2012). Perhaps the most remarkable example is that individuals with positive self-concepts are inclined to seek out and prefer positive feedback, while those with negative self-concepts tend to search for negative feedback (Kwang & Swann, 2010; Swann, Tafarodi, et al., 1992). This phenomenon has lent support to the notion that adding confirming evidence to our self-beliefs might be more important than receiving positive evaluations (Swann Jr. & Brooks, 2012; Swann et al., 1989). Similarly, research indicates that self-discrepant social feedback prompts compensatory responses to protect our self-concept (Swann & Hill, 1982), and that negative feedback exerts a greater detrimental effect on our mood if it is self-incongruent (van Schie et al., 2018). In sum, this research emphasizes that we seek for a stabilized self-concept because it helps understand ourselves, select appropriate interaction partners, and predict our behavior and affect throughout life (Conway, 2005; Steele, 1988, Swan, 2012) Therefore, these proposals cast doubt on the potential of social feedback to trigger changes in our self-concept with the sole motivation of pursuing self-enhancement.

This conflicting evidence leaves, therefore, an important question unanswered. A key aspect to resolve is how we integrate new information from our environment in our self-concept to promote a positively biased but progressively stable self-concept. While research has provided some insights into the role of positivity seeking in self-concept updating, separate research lines would suggest that indiscriminate integration of positive (or negative) feedback might compromise self-concept stability.

To illustrate the importance of integrating insights from these different research lines, consider the following examples. Imagine that we are asked about how sociable we are. Based on our knowledge and experiences, we might be certain that we would never use that trait to describe ourselves. However, we receive social feedback indicating that we are more sociable than we believed to be. A social evaluation suggesting that we should reconsider if we identify as a sociable person would be self-incongruent, but it would also provide the means to see ourselves in a better light. Would that make us align with the social feedback received to increase the positivity of our self-concept? Although accepting positive and incongruent feedback would bring us closer to considering socially desirable traits as our own, it might challenge the stability of our self-concept. Accepting it and shifting our self-evaluations in a feedback-consistent direction, may lead us to internalize a positive trait that has no actual correlate in our behavior. In the opposite scenario, we may recognize ourselves as someone who is sociable. Following a self-evaluation of that characteristic, we receive a social assessment indicating that we should view ourselves as more sociable than we initially thought. Adjusting our self-representation in a feedback-consistent direction not only would increase our self-concept positivity but it would also strengthen the very notion that we are, indeed, sociable. This congruent-like feedback experience is likely to have more chances of being integrated in our self-concept, as it would add confirming (and positive) evidence to an already well-grounded self-representation. Note that the reverse is also true when receiving social evaluations suggesting that we are less sociable than we thought to be. In the case of considering ‘sociable’ as self-descriptive, shifting our self-representations in a feedback-

consistent direction would conflict with the autobiographical evidence we have to support this self-representation (Conway, 2005; Conway et al., 2004). In turn, it would entail a loss of positivity in our self-concept. This might be the perfect scenario to minimize, ignore, or even compensate for the feedback received. In contrast, if we consider sociable as non-self-descriptive, negative feedback would strengthen the notion that we have no evidence to believe we behave as a sociable person. Therefore, it might stabilize our self-concept, provide evidence that we are accurate in our self-judgments (Swan, 2012; Swann Jr. & Brooks, 2012), and distance the possibility of having to internalize a new trait into our self-concept with little to no evidence to draw on. In this case, there is no apparent reason to ignore the feedback received but the fact that being considered sociable is socially desirable (Anderson, 1968). Note that, for negative traits, the same logic can be applied as in the examples above. Taken together, it is plausible to propose that the pursuit of positivity and the need for stability in our self-concept can work in conjunction to shape a self-concept that evolves toward an improved version while preserving its core content. This alignment would be in line with early (Swann et al., 1989) and up-to-date theoretical insights (Mokady & Reggev, 2022).

In the present study, we aimed to orthogonalize and test the effect of both positivity and stability constraints on participants' integration of self-relevant social feedback. In Experiment 1, our objective was to determine the extent to which the initial bias in participants' self-concept would mask the effect of positivity-seeking and stability-seeking motives in updating the self-concept. To further address this issue, in subsequent experiments (Experiments 2 and 3) we isolated the effects of feedback valence (positive vs. negative) and feedback self-congruence (congruent vs. incongruent) and explored their potential as enduring constraints on self-concept updating. Finally, we aimed to investigate the specificity of these effects by testing their generalizability to the updating of representations we hold about others (Experiment 4).

Experiment 1

Introduction

The isolated study of the seek for positivity and the need for self-concept stability has produced apparently conflicting results. Although there are also works testing which one can better explain our social feedback choices (Kwang & Swann, 2010), whether both motives play a role in the way we incorporate social evaluations in our self-concept is a question that remains unanswered. In turn, the lack of consideration of both motives in the study of self-concept updating might lead to an essential conceptual ambiguity. Available evidence has robustly suggested that when updating beliefs about our own traits we tend to integrate positive feedback and discard negative evaluations (Elder et al., 2022; Korn et al., 2012, 2014, 2016). However, given that self-concepts are often characterized by a positive bias (i.e., we tend to perceive ourselves more favorably than unfavorably), the positive feedback participants receive in these studies may primarily be perceived as congruent feedback, whereas negative feedback may be incompatible with participants' current self-representations. Similar concerns have already been pointed out in related literature (Swann Jr. & Brooks, 2012)

In experiment 1, we aimed to reproduce the valence-dependent belief updating found in prior research to exemplify and quantify how participants' initial bias positive bias in their self-evaluations might be masking the effect of positivity and stability constraints on self-concept updating**.** This served as an initial step to develop an experimental paradigm that can effectively isolate the effects of feedback valence and feedback self-congruence on participants' self-concept updating (Experiment 2).

Methods

Participants

Prior to the experiment, we conducted a power analysis using G*Power (Faul et al., 2007) to determine the required sample size. Prior literature in this field has found very large effect

sizes (Korn et al., 2012, 2014, 2016). We assumed a partial eta squared of .1 with a default correlation between measures of .5. Power analysis revealed that for an acceptable power of .8 we needed 20 participants, and we recruited 36 (23 female, M_{age}= 23.22, SD_{age}= 3.94). Participants provided informed consent before their participation. The study protocol was approved by the ethics committee of the University of Barcelona (Institutional Review Board IRB00003099, Comissió de Bioètica de la Universitat de Barcelona).

Procedure

Participants took part in a two-session online experiment on two consecutive days. The first session was administered via Qualtrics (www.qualtrics.com) and the second session via Pavlovia (www.pavlovia.org). The aim of the first session was to create a situation in which participants believed they would receive relevant social feedback during the second session. The second session consisted of performing the experimental task.

First session

In this session, participants were presented with three audio recordings of personality self-descriptions embedded in a Qualtrics questionnaire. Participants were informed that these recordings were randomly selected from other members of the experimental sample. However, the voice clips were made by external collaborators who were not initially informed of the purpose of the study to maintain the authenticity of the recordings. After finishing their voice clips, the collaborators were informed about the study's objectives and provided their consent for their utilization. Each audio recording had a duration of approximately six minutes (min: 5.32, max: 6.73) and the order in which they were presented to the participants was randomly determined. After listening to each recording, participants were asked to evaluate the personality of the speaker by rating a list of 40 adjectives on a scale from 1 (i.e., the adjective does not fit with the description of the person I listened to) to 8 (i.e., the adjective fits perfectly with the person I listened to). Participants were then asked to record themselves describing their personality. They received detailed guidelines of how the audios should be

made and they were told to use the three recordings presented previously as examples. The guideline consisted of 10 items randomly drawn from the HEXACO personality questionnaire (https://hexaco.org/) (e.g., “I feel reasonably satisfied with myself overall”, “I rarely express my opinion in social meetings”). Participants were instructed to speak for at least 30 to 45 seconds about each statement, offering their degree of agreement with them and explaining the reason for their answer with examples or anecdotes. Once the recording was completed, participants attached it to the online questionnaire. Finally, participants were requested to complete the Beck Depression Inventory (BDI-II), which 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, one participant scored 22 and was excluded from further analysis. Participants who had missed more than 15% of the trials in any of the two experimental tasks (see below) were also excluded from the sample. Two participants met these later criteria and were excluded from further analysis.

Second session

In the second session, participants performed two consecutive experimental tasks. First, participants underwent a sentence verification type task (Figure 1 A), which has previously been used to study personal semantic memory and self-representations (Renoult et al., 2012). This task consists of accepting and rejecting sentences by means of a dichotomous response, typically "Yes" vs "No". In this task, participants were asked to decide whether the adjectives presented in each sentence represented themselves or not (e.g., "I am friendly"). The trial structure consisted of a fixation cross (500 ms.) followed by the sentence "I am" (1 second). After another fixation cross (500 ms) an adjective appeared on the screen (e.g., "friendly", 3 seconds). At that moment, they had 3 seconds to provide their responses using the left (Yes) and right (No) arrows on their computer keyboards.

Next, participants responded to a belief updating task that has been previously used to study the impact of positive and negative feedback on beliefs about one's own personality traits

(Elder et al., 2022; Korn et al., 2012, 2014). In each trial, participants were presented with 1 of 40 trait adjectives together with the prompt “how do you see yourself?” and had to think about how much that trait applied to themselves and rate it on a Likert scale from 1 (this trait does not describe me at all) to 8 (this trait describes me perfectly) (3 seconds). After another fixation cross (500ms) participants saw what they believed to be the mean rating that three other participants provided about them on that trait after listening to their audio recording (2 s.). That rating, which served as a feedback rating, was a number ranging from 1.0 to 8.0 in steps of 0.5. As in prior studies, feedback ratings were pseudo-randomly generated by the program during the experiment to produce a balanced number of trials in which participants received positive and negative feedback. After repeating this process for all adjectives, participants rated themselves again on all traits (reassessment). Once both tasks were completed, participants were asked to recall the feedback ratings (Likert scale from 1 to 8) they had received for each trait in a separate questionnaire.

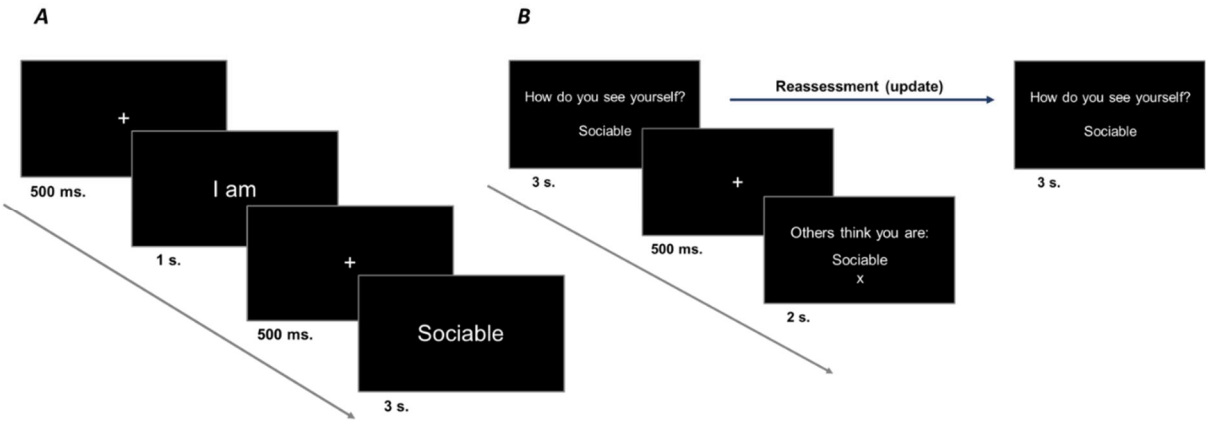

Figure 1. Experiment outline. Sentence verification task (A). During the task, participants were tasked with determining whether the adjectives presented in each sentence described themselves or not (e.g.,

"I am friendly"). The trial structure involved a sequence of events, starting with a fixation cross lasting 500 ms., followed by the sentence "I am" displayed for 1 second. After another 500 ms. fixation cross, an adjective appeared on the screen (e.g., "friendly") and remained visible for 3 seconds. Participants had 3 seconds to respond by using the left arrow (for "Yes") or the right arrow (for "No") on their computer keyboards. Belief updating task (B). In each trial, participants were presented with 1 of 40 trait adjectives together with the prompt “how do you see yourself?” and had to think about how much that trait applied to themselves and rate it on a Likert scale from 1 (this trait does not describe me at all) to 8 (this trait describes me perfectly) (3 seconds). After another fixation cross (500 ms.) participants saw what they believed to be the mean rating that three other participants provided about them on that trait after listening to their audio recording (2 s.). That rating, which served as a feedback rating, was a number ranging from 1.0 to 8.0 in steps of 0.5. As in prior studies (Korn et al., 2012), feedback ratings were pseudo-randomly generated by the program during the experiment to produce a balanced number of trials in which participants received positive and negative feedback. After repeating this process for all adjectives, participants rated themselves again on all traits (re-evaluation).

Main measures

The main measures extracted from the experimental tasks were participants’ categorical responses (Yes vs. No), pre- and post-feedback self-assessments (8 points Likert scale), valence of the feedback received (positive vs. negative), and feedback rating (8 points Likert scale, in steps of .5). Pre- and post-feedback self-assessments were used to compute the dependent variable of interest of the study, namely, update scores. Update scores represent the degree of change of initial beliefs after receiving social feedback. We computed this variable so that it could be interpreted as the degree of feedback-consistent update (i.e., the degree to which a belief changes in the direction that the feedback suggests) (Korn et al., 2012). Specifically, this measure was computed as follows: when the feedback received was lower than the self-assessment, update scores were computed as self-assessment(pre) - self-assessment(post). When the feedback received was higher than the self-assessment, update scores were equal to self-assessment(post) - self-assessment(pre). Note, that we allowed

negative values for this variable, which denote that the participant has changed their self-representations in the opposite direction to what the feedback suggested. Feedback valence was determined as follows: if a positive adjective received a feedback rating lower than the self-rating or a negative adjective received a higher feedback rating than the self-rating, the feedback was labelled as negative. If the feedback received for a positive adjective was higher than the self-rating or the feedback for a negative adjective was lower than the self-rating, feedback was labelled as positive. Categorical responses from the sentence verification task were used in combination with feedback valence to generate a measure of feedback self-congruence. Feedback self-congruence was operationalized as a two-level factor (self-congruent vs self-incongruent) that aimed to capture the extent to which the feedback received had the potential to either reinforce or conflict with participants' self-representations. To illustrate this idea, imagine a participant who in the sentence verification task decides that the trait "sociable" does not apply to them. Subsequently, in the belief updating task when they have to provide a self-rating on the same trait, they select a "3" (out of 8). In the case of receiving a score lower than 3 as (negative feedback), the feedback received would be supporting their notion that they are not a sociable person (self-congruent feedback). In contrast, if they receive a score higher than 3 (positive feedback), the feedback received would be suggesting that they should reconsider the notion that they are not a sociable person (self-incongruent feedback). As in most studies in related literature, we also computed a covariate aimed to capture the degree of miss-match between participants' self-ratings and feedback ratings, namely, feedback discrepancy. This measure was computed for each participant as the average absolute difference between self-ratings and feedback ratings. We introduced another measure to provide additional control to the analysis. This measure was intended to capture how much space within the scale participants had available for updating. We named this measure update space. The update space was computed as the average difference between participants' self-assessments(pre) and the limit of the scale (8-point Likert scale in our case) taking into account the direction suggested by the feedback. That is, if self-

assessment(pre) < feedback, then update space = 8 - self-assessment(pre). If self-assessment(pre) > feedback, then update space = self-assessment(pre) - 1.

Stimuli

For this experiment, we randomly selected 20 positive and 20 negative adjectives from prior studies (Korn et al., 2012, 2014), which come from a widely used list of personality attributes (Anderson, 1968). We asked a separate sample of participants (n=42) to provide observability ratings on each adjective. Specifically, we asked participants to decide the extent to which each adjective would be discerned from a given individual by listening a 6-minute personality description. We provided participants the specific guidelines those individuals would use to describe themselves. Participants provided those ratings using a Likert scale ranging from 1 (not observable at all) to 8 (very observable). We selected only those adjectives with an average observability rating above 4.5. The final sample of stimuli consisted of a list of 31 adjectives describing character traits (16 positive and 15 negative). To equate the number of adjectives between valence categories we selected one negative adjective to complete the list based on its closeness to our inclusion criterion. Selected stimuli were classified as positive or negative according to their previously reported average desirability ratings (Anderson, 1968), and prior classifications (Korn et al., 2012, 2014, 2016). To check if this classification remained stable in our sample, we asked participants to provide desirability ratings for all the adjectives in the list. We then compared participants' average ratings of positive vs. negative adjectives by means of a paired t-test. The Paired t-test testing the difference between desirability ratings on positive vs. negative adjectives showed that positive adjectives received significantly higher desirability ratings than negative adjectives (difference = 2.068, 95% CI [2.849, 3.286], $t(33) = 28.578$, $p < .001$; $d = 4.974$, 95% CI [3.338, 5.591]). In addition, we tested whether the desirability ratings of positive adjectives were above the mid-point scale (4.5) by means of a one-tailed one-sample t-test (against mu = 4.5). Finally, we tested if the desirability ratings of negative adjectives were below the mid-point scale (4.5). Desirability ratings of adjectives classified as positive showed to be significantly above 4.5 ($t(33) = 29.946$,

p < .001; Cohen's d = 5.212, 95% CI[4.105, Inf]) and desirability ratings of adjectives classified as negative showed to be significantly below 4.5 t(33) = -13.993, p < .001; d = -2.435, 95% CI[-Inf, -1.856]). For studies 2, 3, and 4 we selected and screened 120 adjectives with a separate sample (n = 82) according to different criteria (see, Stimuli, Study 2).

Results

Following prior research, we first assessed whether participants updated their self-representations more in response to positive than negative feedback. To that end, we conducted a repeated measures Analysis of Variance (rmANOVA) with average update scores as the dependent variable, Feedback valence (positive vs. negative) as a within-participants factor, and feedback discrepancy and update space as covariates. Results showed that participants tended to update more their beliefs in response to positive than to negative feedback (positive feedback: M= .376 SE= .109 95% CI[.157, .549]; negative feedback: M= -.195 SE= .109 95% CI[-.414,.023], $F(1,30) = 12.004$, $p = .001$, $\eta 2 = .285$, 90% CI[.076, .461]). These results replicated the valence-dependent belief updating effect described in previous literature.

In line with the main goal of this study, we quantified how participants' initial self-evaluations masked the effect of feedback valence and feedback self-congruence on participants' self-concept updating. We assumed participants self-concepts to be positively biased, i.e., they would make more positive than negative categorical decisions about their attributes. This initial positive bias creates a scenario where most positive feedback overlaps with feedback that aligns with the individual's self-views (self-congruent feedback), and likewise, most negative feedback overlaps with feedback that conflicts with their current self-concept (self-incongruent feedback). This potential overlap would obscure the distinct effects of feedback valence and feedback self-congruence. In the most extreme scenario, a participant who only makes positive categorical self-judgments would find positive feedback 100% self-congruent (confirming their notion that all positive traits are self-descriptive, and all

negative traits are not). In contrast, they would find all negative feedback incongruent, as it would challenge their self-concept in every trial. To quantify this overlap, we first tried to predict participants' categorical judgments (yes vs. no) using adjective valence (positive vs. negative) as a predictor, by means of a mixed-effects logistic regression (participants' ID as a random effect). The results of this analysis showed that adjective valence strongly predicted participants' categorical self-judgments (Marginal R2: .383, Conditional R2: .391). Specifically, we found that it was 17.727 times more likely to choose a positive than a negative adjective as self-descriptive (Odds ratio[positive]: 17.727, SE = 3.263, 95% CI[12.358, 25.430], z = 15.618, p <.001).

Next, we aimed to quantify to which extent a positive feedback trial would be also classified as a congruent feedback trial, which provides a measure of the overlap between conditions. The results of this analysis suggested that a positive feedback trial was 11.049 times more likely to be classified also as a congruent than a negative feedback trial (Odds ratio[positive feedback]: 11.049, SE = 1.634, 95% CI[8.268, 14.764], z = 16.242, p <.001). Our results, therefore, suggested that the belief updating bias attributed to the effect of feedback valence could be similarly ascribed to feedback's congruence with participants' initial self-concept.

Finally, we compared updating scores averaging them across feedback self-congruence categories (self-congruent vs. self-incongruent) instead of across feedback valence conditions. Analogously to the results obtained with feedback valence, the results suggested that participants tended to update more their beliefs in response to self-congruent feedback than to self-incongruent feedback (self-congruent feedback: M= .617 SE= .175 95% CI[.266, .968]; self-incongruent feedback: M= -.456 SE= .175 95% CI[-.807,.105], $F(1,30) = 7.484$, $p = .011$, $\eta 2 = .199$, 90% CI[.028,.382]).

Discussion

In experiment 1, we showed that to appropriately evaluate the possibility that our self-representations are updated in a valence-dependent manner, we should first disambiguate the effect of feedback valence from feedback self-congruence. To do so, we should first address participants' positive bias in self-concept, as it produces an overlap between positive-congruent and negative-incongruent feedback trials.

The discussion regarding the degree to which individuals emphasize self-concept positivity as opposed to stability has been a longstanding topic in the literature (Kwang & Swann, 2010). New research has offered a more nuanced perspective on positivity constraints on self-concept updating by examining how individuals update their self-concept in response to self-relevant feedback. This research has shown that we not only strive to receive positive feedback and seek out those who give it to us (Alicke & Sedikides, 2009; Hepper et al., 2010; Pyszczynski et al., 1985) but in fact, we asymmetrically incorporate positive vs. negative self-relevant feedback to increase our self-concept positivity (Elder et al., 2022; Korn et al., 2012, 2014, 2016). Although this research has provided robust results and large effect sizes it mainly focuses on studying how positive feedback selectively induces self-concept update (but see, Elder et al., 2022). This approach overlooks the potential of that feedback to reinforce or conflict the current self-concept, thereby enhancing or challenging self-concept stability. Our results suggest that existing evidence cannot conclusively attribute the observed biases in self-concept updating solely to the pursue of self-concept positivity. Indeed, they might be similarly explained by our need to maintain a stable self-view.

The orthogonalization of feedback self-congruence and feedback valence would allow us to study at the same time whether motives for positivity and stability drive the integration of social evaluative feedback in self-concept representations.

Experiment 2

Introduction

The findings from Experiment 1 highlighted the importance of distinguishing between positivity-seeking and stability-seeking influences when studying the incorporation of feedback into the self-concept. To accurately disentangle these effects, it is necessary to control for potential positive biases in participants' self-concept. The majority of the population tends to have a positively biased self-concept (Alicke & Sedikides, 2009; Taylor et al., 1988). However, it is common for individuals to recognize some negative aspects of themselves and acknowledge the absence of certain socially desirable traits they may wish to possess (Baranski et al., 2021; Higgins, 1987; Steele, 1988). We propose that an effective way to control for the effect of the positive bias in the self-concept might be to sample individuals' beliefs about their traits. Our proposal involves utilizing a non-proportional stratified random sampling method to gather participants' positive and negative decisions concerning their self-concept. This approach would ensure that for every participant a balanced set of positive and negative decisions is available. Importantly, this approach could also be used to control the effect of both positively and negatively biased self-concepts.

As outlined before, an exclusive focus on either seeking positivity or maintaining stability in our self-concept could be suboptimal for our well-being. Alternatively, both motives might have room to influence how we deal with self-relevant feedback (Mokady & Reggev, 2022; Swann et al., 1989). Our drive for stability in our self-representations may result in the preferential incorporation of self-congruent feedback. Its integration would add consistent evidence to an already well-grounded self-concept (Conway, 2005; Rathbone et al., 2008), which would be neutral in terms of positivity gains (it would confirm both our positive and negative aspects, or lack thereof). Gaining self-concept stability could in itself be beneficial for well-being (Campbell, 1990), but it could compromise progress towards a more positive self-concept, as it might also imply adding confirming evidence on negative traits or our lack of desirable traits.

If that is the case, we hypothesize that this could be solved by biasing the integration of self-congruent feedback in favor of that evaluations that besides being self-congruent, are positive in nature. In turn, a drive for positivity may lead to an enhanced integration of positive feedback leading to see ourselves in a better light. However, the indiscriminate integration of positive feedback over negative feedback could be detrimental to the stability of the self-concept. Specifically, it might involve ceasing to identify ourselves with negative attributes strongly supported by our autobiographical knowledge or starting to believe that we possess traits that do not reflect our behavioural patterns. This might lead to increased uncertainty about ourselves, our abilities, goals, or feelings (Conway, 2005; Conway et al., 2004; Kim & Chiu, 2011; Swann, 2012; Swann Jr. & Brooks, 2012). Note that accepting negative self-congruent feedback to the same extent as negative self-incongruent feedback might also be suboptimal. Although assuming that negative social evaluations would be integrated to a lesser extent, the integration of congruent negative feedback would reinforce our certainty about our self-concept, while integration of self-incongruent feedback would bring the double penalty of loss of positivity and stability.

We propose that the optimal pattern of integration of self-relevant feedback would likely be one that balances both self-concept enhancement and self-concept maintenance. In experiment 2, we aimed to test the effect of both positivity and stability constraints on participants' integration of self-relevant social feedback. We hypothesized that participants would incorporate more self-congruent than self-incongruent feedback into their self-representations as well as more positive than negative feedback. Here, we also aimed to explore whether feedback-induced changes in self-representations would remain one day after the experiment. This test may provide further understanding of how self-concept representations are shaped and maintained, along with their potential variability depending on the type of feedback received.

Methods

Participants

For this experiment, we recruited 45 participants (undergrate students, 29 female, M_{age}= 22.19, SD_{age}= 2.11). Required sample size was obtained by another power analysis, which determined that to detect an effect of η^2 = .05 we would need at least 28 participants. Although prior studies have found very large effect sizes, we decided to reduce the effect size parameter because we expected some of the explained variance attributed to feedback valence to be captured by feedback self-congruence. Three participants were excluded from the sample based on their BDI-II scores (participant 1: 21, participant 2: 23, participant 3: 23). One participant was excluded from the sample based on their number of missing responses [>15%, (participant percentage of missing responses = 46.87%)]. The final sample was composed of 41 participants (28 female M_{age}= 22.44, SD_{age}= 2.32).

Procedure

Participants took part in a three-session experiment, on three consecutive days. The objective of the initial session was to set up a scenario where participants anticipated receiving relevant social feedback during the subsequent session. The second session consisted of performing the experimental tasks (see, Experiment 1). The aim of the third session was to obtain a follow-up measure of the effects studied in session 2. All participants were informed about the experimental manipulation right after the conclusion of session 3.

As outline above, to effectively investigate how the pursuit of maximizing positivity or stability influences the incorporation of information into the self-concept, it is crucial to distinguish and analyze their individual effects. To accomplish this goal, the overlap that our positively biased self-concept posits between positive and self-congruent feedback as well as between negative and self-incongruent feedback needs to be experimentally controlled. Here we introduced a novel approach to that end, namely, a non-proportional stratified random sampling of

participants' positive and negative decisions about the self. This process allows an unbiased part of the participants' self-concept to be subjected to the belief updating task. The procedure was introduced in the sentence verification task (Fig 1 A). As outlined in study 1, this task involves participants making positive (rejecting negative traits or accepting positive traits) and negative (rejecting positive traits or accepting negative traits) self-relevant decisions about various traits. We used an extended list of adjectives (n= 95, see below, *Stimuli*) to increase the number of decisions participants would make about themselves. The non-proportional stratified random sampling consists of a random selection of the same percentage of elements in all strata (here positive and negative decisions) to achieve a balanced sample. In the sentence verification task, participants judged the 95 adjectives in random order and provided confidence judgments for all decisions. From all the (mostly positive) decisions made, a minimum number of them per stratum was extracted. That is, the same number of positive and negative decisions were randomly drawn from their respective populations. Based on the data provided by study 1, we estimated that the percentage of positive decisions participants would make would as much be between 80% and 82% of the total (study 1 upper bound 99% CI = 81.5%) which would imply making between 77 and 79 positive decisions out of 96 adjectives in the most extreme cases. To obtain balanced conditions (positive and congruent feedback, positive and incongruent feedback, negative and congruent feedback, and negative and incongruent feedback) half of the positive decisions should receive negative feedback and the other half negative feedback, and the same for negative decisions. Therefore, we set at 16(/96) the minimum number of negative decisions necessary for a participant to be included in the analysis. One participant did not meet this inclusion criterion and was excluded from further analysis (effective sample size: n = 40). Once this process was completed and a balanced set of positive and negative decisions was reached, participants underwent the belief updating task.

Participants were also tested one day after the completion of both tasks to test the potential long-lasting effects of feedback valence and feedback congruence on participants' self-

representations. In this session, participants provided again categorical decisions about their own traits. Judgments only included those participants' decisions that were filtered in session two by the non-proportional stratified random sampling. Next, participants provided self-ratings (Likert scale from 1 to 8) on the same traits, as in the first part of the belief updating task.

Stimuli

The non-proportional stratified random sampling of participants' positive and negative decisions required us to extend the list of adjectives used in the pilot study. To this end, we selected an initial sample of 120 adjectives drawn from previous studies (Anderson, 1968; Dumas et al., 2002; Korn et al., 2012, 2016) on the basis of their valence (60 positive and 60 negatives). To further refine the list of adjectives, we aimed to obtain measures of valence and familiarity from a separate sample of participants with similar characteristics (age and educational level) to the one that would carry out our experiment. We recruited a sample of 82 university students (57 females, Age: M= 21.57, SD= 1.64). Participants judged adjectives' valence and familiarity using a Likert scale ranging from 1 to 8. Since our experiments involved participants evaluating other people on the basis of personality descriptions and believing that other people would evaluate them, we also wanted to select those adjectives that were judged to be 'most observable' in this context. Thus, as in the separate sample we recruited for the same purpose in Experiment 1, we asked participants to assess the extent to which they considered each adjective to be 'observable' by listening to a recording of a 6-minute personality description. We provided each participant with the guidelines that participants in our experiment were to use to make their recordings (i.e., items taken from personality questionnaires). To filter the final sample of adjectives, we selected as positive adjectives those with an average score equal to or higher than 5 (on a scale from 1 to 8), and as negative adjectives those with a score equal to or lower than 4. In addition to this criterion, only those adjectives with an average familiarity and observability score above 4.5 were selected. The final set of stimuli consisted of 47 positive and 48 negative adjectives.

Results

To test the effect of both feedback valence and feedback self-congruence on participants' belief updating we conducted a rmANOVA with average updating scores as the dependent variable, Feedback valence, Feedback self-congruence and their two-way interaction as within-participants factors, and Feedback discrepancy and Update space as covariates. Estimated marginal means (Lenth, 2022) are reported. Results showed that participants tended to update significantly more their self-representations in response to Self-congruent feedback than in response to Self-incongruent feedback (Self-congruent: M= .472 SE = .133 95% CI[.210, .733], Self-incongruent M= -.131 SE = .133 95% CI[-.393, -.131], F(1, 37) = 14.081, p <.001, η2 = .28, 90% CI[.091, .442]). Results also showed a non-significant main effect of Feedback valence (Positive feedback: M= .112 SE = .061 95% CI[.107, .348], Negative feedback M= .228 SE = .061 95% CI[-.008, .233], F(1, 37) = .276, p = .603, η^2 = .007, 90% CI[0, .051]) and a non-significant Feedback Self-congruence x Feedback valence interaction (F(1, 37) = .048, p = .828, η^2 = .001, 90% CI[0, .018]) (Figure 2, A).

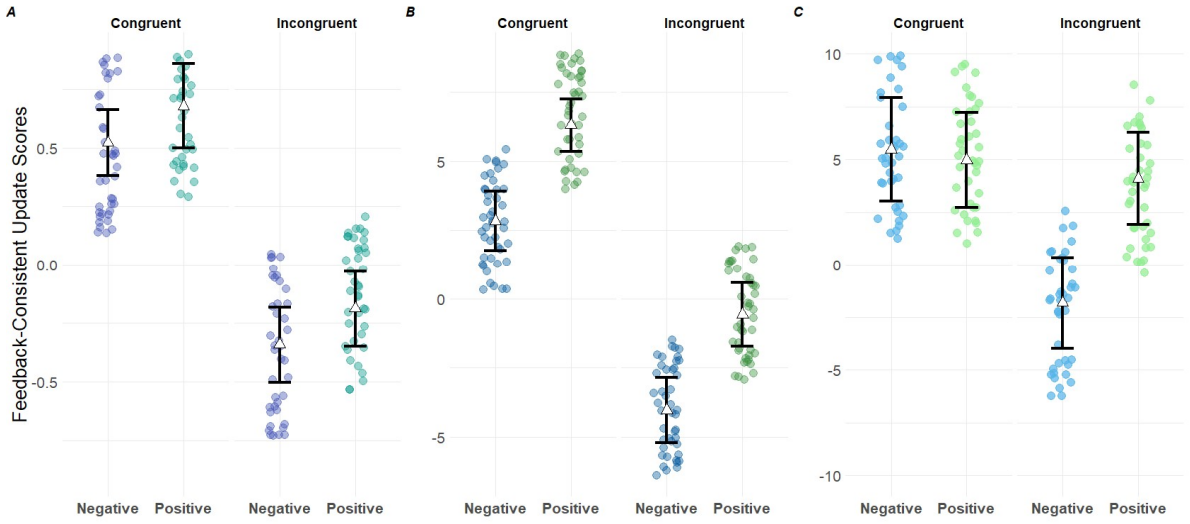

Figure 2. Differences in Feedback-consistent belief update scores between feedback self-congruence (facet) and feedback valence conditions (colors) in experiments 2 (A), 3 (B) and 4 (C). Triangles represent the estimated marginal means of each feedback condition (Lenth, 2022), the 95% confidence interval is indicated by the ends of the vertical error bar.

Next, we aimed to test whether the observed effects remained 24h hours after the experimental session. 4 (10%) participants did not participate in the follow-up test. We conducted a rmANOVA with update scores as the dependent variable, feedback self-congruence, feedback valence, time (session 2, follow-up) and their three-way interaction as within-participants factors, and feedback discrepancy and update space as covariates. Results showed a significant main effect of feedback self-congruence (self-congruent: M= .548 SE = .093, 95% CI[.365, .732], self-incongruent M= -.288 SD = .093, 95% CI[-.472, -.105], $F(1, 33) = 14.069$, $p < .001$, $\eta^2 = .298$, 90% CI[.094, .466]). No significant effects were found for feedback valence (positive feedback: M= .197 SE = .055, 95% CI[.087, .308], negative feedback M= -.062 SE = .055, 95% CI[-.047 , .173], $F(1, 33) = .851$, $p = .363$, $\eta^2 = .025$, 90% CI[0, .111]), time (session 2: M= .147 SE = .047, 95% CI[.053, .241], follow-up M= .113 SE = .047, 95% CI[.018, .207], $F(1, 35) = 1.451$, $p = .236$, $\eta^2 = .039$, 90% CI[0, .140]), Feedback self-congruence x feedback valence interaction ($F(1, 33) = .011$, $p = .917$, $\eta^2 = .<001$, 90% CI[0, .001]), feedback self-congruence x time interaction ($F(1, 35) = .849$, $p = .363$, $\eta^2 = .023$, 90% CI[0, .103]), feedback valence x time interaction ($F(1, 35) = 2.344$, $p = .135$, $\eta^2 = .062$, 90% CI[0, .184]), or feedback self-congruence x feedback valence x time interaction ($F(1, 35) = .109$, $p = .744$, $\eta^2 = .003$, 90% CI[0, .033]).

As suggested by prior studies (Garcia-Arch et al., 2022; Korn et al., 2012), we explored the distribution of feedback discrepancies among the different experimental conditions. We conducted a rmANOVA with average feedback discrepancy as the dependent variable, and feedback self-congruence, feedback valence, and feedback self-congruence x feedback

valence interaction as within-participants factors. We found a significant main effect of feedback valence (F(1, 39) = 18.959, p <.001, η2 = .33, 90% CI[.134, .483]), a significant main effect of feedback self-congruence (F(1, 39) = 456.901, p <.001, η2 = .92, 90% CI[.873, .940]), and a significant feedback self-congruence x feedback valence interaction (F(1, 39) = 7.054, p = .011, η2 = .15, 90% CI[.018, .308]). Post-hoc contrasts (Estimated Marginal Means, Tukey's p-value adjustment) revealed that all self-incongruent conditions received significantly larger feedback discrepancies than self-congruent conditions. Results also showed that the self-congruent + negative feedback condition was associated with higher feedback discrepancies than the self-congruent + positive feedback condition (p <. 05). These results suggest that our experimental conditions were strongly biased in terms of the feedback discrepancies they received.

Discussion

In experiment 2, we aimed to examine how both the motivation for a positive self-concept and the motivation for a stable self-concept influence participants' incorporation of self-relevant social feedback. We employed a method that allowed us to control the effect of participants' positively biased self-concept, and to generate balanced experimental conditions in which participants received self-congruent, self-incongruent, positive, and negative social evaluations. We hypothesized that participants would incorporate more self-congruent than self-incongruent feedback into their self-representations as well as more positive than negative feedback. Our results showed that participants integrated substantially more self-congruent than self-incongruent feedback into their self-representations by adjusting their self-ratings in a self-congruent feedback-consistent direction. This effect was stable one day after the experimental session. These findings are in line with models that suggest that the main motivation for our self-concept is to be socially verified, which allows us to preserve identity stability and reinforces our perceived accuracy in our self-judgments (Swann, 1982; Swann, et al., 1992) but extend them by indicating that the pursuit of stability constrains the capacity of social evaluative inputs to alter self-representations.

Interestingly, feedback valence showed no effect in participants' updating of self-representations. This finding challenges prior research that showed a strong tendency to integrate more positive than negative social evaluative feedback into our self-representations (Elder et al., 2022; Korn et al., 2012, 2014, 2016). The lack of bias towards positive feedback suggests that new self-congruent information is readily integrated into the self-concept regardless of its valence. These results may be interpreted as indicating that individuals strive to stabilize their self-concepts by incorporating new confirming evidence that allows them to reinforce their self-views, even if their self-concepts include negative features or lack socially desirable attributes. However, upon further examination of the data, it was observed that a key variable influencing belief updating, namely, feedback discrepancy (Korn et al., 2012), was not evenly distributed across the different feedback categories. Although its effects were statistically controlled for in our analysis, it is possible that the systematic differences in feedback discrepancies found across feedback conditions might have biased participants' behavior. Receiving feedback ratings more different from their own self-assessments in the self-incongruent conditions might lead participants to perceive self-incongruent feedback even more self-incongruent, leading to its detrimental integration. However, the opposite interpretation is also possible, i.e., these larger feedback discrepancies might have heightened its integration and reduced the incorporation of the self-congruent feedback, which may have been integrated to a greater extent if the feedback discrepancies had been evenly distributed across all conditions. In turn, the lack of a feedback valence effect could be partly explained by the fact that participants received larger feedback discrepancies in the self-congruent negative feedback than in the self-congruent positive feedback condition.

Any of these or other possible interpretations suggest that beyond statistical control, feedback discrepancies should be controlled experimentally. In fact, as in previous studies (see, *procedure*, Experiment 1), feedback discrepancies were pseudo-randomly generated so that half the time they resulted in positive feedback and the other half in negative feedback. This manipulation generates conditions with similar feedback discrepancies between

conditions when these are split into positive and negative feedback (Garcia-Arch et al., 2022; Korn et al., 2012). However, this manipulation does not take into account self-congruence. In the self-congruent feedback conditions, participants receive feedback that is 'confirmatory'. This means that for participants who judge the trait "Sociable" as non-representative and assign themselves a 3 in their self-rating, the feedback they will receive can only be lower than 3. In this same example, if this trait were in the self-incongruent condition, the feedback received could take any value above 3 and up to 8. The randomization of feedback discrepancies is consequently biased between the two conditions because its values are sampled from unequal value ranges. We addressed this issue in study 3 along with additional considerations.

Experiment 3

Introduction

Findings from experiment 2 revealed that only social feedback that confirms our self-concept seems to be integrated into self-representations. However, upon closer analysis of the data, we observed a notable imbalance in feedback discrepancies among the different conditions, which might have influenced the results. Thus, our aim was to make methodological adjustments to enhance the robustness of the evaluation of the main hypotheses in Experiment 2.

Methods

Participants

For this experiment, we recruited 52 participants. Two participants were excluded from the sample based on their BDI-II scores (score participant 1: 32, score participant 2: 24). One participant was excluded from the sample based on their number of missing responses [>15% (% missing responses: 90%)]. As in Experiment 2, we also excluded from the sample those participants that did not make a minimum number of negative decisions about their self-

concept. This procedure excluded only one participant in Experiment 2, where the minimum number of negative decisions was set to 16/96. Given the low exclusion rate in Experiment 2, we decided to increase the total number of trials of the experiment and the required number negative decisions accordingly (20/96). Four participants did not reach the minimum number of negative decisions required and were excluded from the sample. The final sample was composed of 45 participants (32 female M_{age}= 21.76, SD_{age}= 1.39).

Procedure

The procedure employed in this study was the same as in Experiment 2, with two modifications. First, to avoid asymmetries in the distribution of feedback discrepancies across feedback conditions, we restricted the range of values among which feedback was sampled to be the same in all conditions. We adjusted this range based on the condition with lower average feedback discrepancies according to study 2. Feedback values were computed by adding or subtracting values from 1 to 15 to participants self-ratings (see, *Methods*, Study 1, for details). Second, and in line with the same aim, we changed the 8 points Likert scale to a scale with a wider range of values (1 to 100), which has been also used in belief updating paradigms (Sharot & Garrett, 2016). The main reason to consider the 8 points Likert scale problematic was that it generates a range restriction in feedback values and possible updates. The limited 8-point range for self-ratings used in prior studies (Korn et al., 2012, 2014, 2016), as well as in our experiments 1 and 2, has the limitation that in some scenarios it may influence the potential updating of participants' beliefs. To exemplify one of its limitations, imagine that a participant provides a self-rating of 5 for the adjective 'sociable'. If this participant receives a 7 as a feedback rating, they could update their belief in a feedback-consistent direction by moving it up to the feedback value offered. This scenario already entails a problem, as it would only allow counting as updates those belief changes that represent a percentual change of at least 12.5% (increments or decrements of 1 over the total scale), which could imply a loss of belief updating sensitivity. A more problematic case might arise when the feedback received only implies a discrepancy of (+/-) .5, or (+/-) 1. In these cases, belief updating could not occur

within the range between the initial belief and the feedback received, since self-assessments are limited by the 8-point Likert scale. Changing to a wider (1 to 100) scale has the potential to solve both problems by providing sufficient range of values for both self and feedback ratings.

Results

In this study, we aimed to solve the asymmetry in feedback discrepancies across feedback conditions found in experiment 2 to provide a more reliable test for our hypotheses. To check if our methodological adjustments managed to cancel out the asymmetry in feedback discrepancies across feedback conditions, we conducted a rmANOVA with average feedback discrepancy as the dependent variable, and feedback self-congruence and feedback valence as within-participants factors. Results showed no significant feedback discrepancy asymmetries across feedback conditions (feedback self-congruence $F(1, 44) = .182$, $p = .699$, $\eta2 = .003$, 90% CI[0, .030], feedback valence $F(1, 44) = .011$, $p = .919$, $\eta2 = <.001$, 90% CI[0, .007], feedback self-congruence X feedback valence $F(1, 44) = .066$, $p = .797$, $\eta2 = .001$, 90% CI[0, .016]).

After testing the effectiveness of the experimental control on feedback discrepancies, we set out to reproduce the main analysis carried out in Experiment 2. We conducted a rmANOVA with average update scores as the dependent variable, feedback valence, feedback self-congruence and their two-way interaction as within-participants factors, and feedback discrepancy and update space as covariates. Results showed a significant and large main effect of feedback self-congruence ($F(1, 42) = 16.166$, $p = <.001$, $\eta^2 = .277$, 90% CI[.098, .431]), a significant and medium effect of feedback valence ($F(1, 42) = 4.517$, $p = .039$, $\eta^2 = .097$, 90% CI[.002, .231]) and no significant feedback self-congruence x feedback valence interaction ($F(1, 42) = 2.968$, $p = .092$, $\eta^2 = .065$, 90% CI[0, .179]). Next, we aimed to conduct a more fine-grained analysis by means of linear mixed-effects models (LMMs). This approach allows to capture and account for individual differences in the effects tested, provide

interpretable estimates for any term of the model, compute proper marginal effects and post-hoc tests, and incorporate additional random effects in the covariance structure of the model tested (Baayen et al., 2008; Barr et al., 2013; Brown, 2021). To obtain the best LMM we constructed alternative models that varied in their inclusion of random intercepts and slopes and compared them by means of the Bayesian Information Criteria (BIC), which penalizes model complexity (Schwarz, 1978). P-values were determined by Satterthwaite's approximation of degrees of freedom (Kuznetsova et al., 2017). For model construction, we started with the LMM version of our rmANOVA model, which included fixed effects for feedback self-congruence, feedback valence, feedback discrepancy and update space, and participants'ID as a grouping factor. We subsequently tested if this model could be improved by sequentially adding different combinations of: partially-crossed random effects (Adjectives and Participants' ID), the interaction between feedback self-congruence and feedback valence, and random slopes for feedback self-congruence and feedback valence. Maximal random effects structures were kept when supported by the data and model convergence (Barr et al., 2013). Among all LMMs, only one outperformed the starting model. This model included fixed effects for feedback self-congruence, feedback valence, feedback discrepancy and update space, participants'ID as a grouping factor and a random slope for feedback valence (Maringal R^2 = .166, Conditional R^2 = .204). LMM converged with the results obtained by the rmANOVA. Results showed that participants tended to update significantly more their self-representations in response to self-congruent feedback than in response to self-incongruent feedback ($\beta_{Self\text{-}congruent}$ = 6.426, SE = .698, 95% CI[5.056, 7.797], t(1701.314) = 9.196, p <.001) and in response to positive (vs negative) feedback ($\beta_{positive}$ = 3.586, SE = .731, 95% CI[2.153, 5.021], t(46.186) = 4.906, p <.001) (Figure 2, B). Results also showed no significant relationship between feedback discrepancy and update scores (β = .031, SE = .072, 95% CI[-.172, .112], t(1732.979) = -.419, p =.675) and a positive and significant relationship between update space an update scores (β = .213, SE = .012, 95% CI[.189, .236], t(1722.959) = 17.439, p <.001). To further explore differences in feedback-consistent belief updating across feedback conditions we computed Sidak adjusted confidence intervals (Cis) on the

marginal means (Lenth, 2022). The analysis showed that participants integrated social evaluative feedback in their self-representations (adjusted CIs did not include 0) in both self-congruent feedback conditions (positive feedback: M = 6.117, SE = .627, 95% CI[4.528, 7.705], negative feedback: M = 2.531, SE = .553, 95% CI[1.123, 3.937]). Results also revealed that self-incongruent and positive feedback did not affect participants self-representations (M = -.309, SE = .578, 95% CI[-1.781, 1.162]) while self-incongruent and negative feedback tended to be rejected, i.e., participants tended to update their self-representations in the opposite direction (M = -3.895, SE = .608, 95% CI[-5.432, -2.359]).

Next, we aimed to test whether the observed effects remained one day after the experimental session. 5 (11%) participants did not participate in the follow-up test. We conducted a rmANOVA with update scores as the dependent variable, feedback self-congruence, feedback valence, time (session 2, follow-up) and their three-way interaction as within-participants factors, and feedback discrepancy and update space as covariates. Results showed a significant main effect of feedback self-congruence ($F(1, 37) = 14.708$, $p < .001$, $\eta^2 = .284$, 90% CI[.093, .446]) and a significant main effect of feedback valence ($F(1, 37) = 5.174$, $p = .028$, $\eta^2 = .122$, 90% CI[.006, .275]). No significant main effects were found for time ($F(1, 39) = .056$, $p = .814$, $\eta^2 = .001$, 90% CI[0, .017]), feedback self-congruence x feedback valence interaction ($F(1, 37) = .920$, $p = .343$, $\eta^2 = .024$, 90% CI[0, .103]), feedback self-congruence x time interaction ($F(1, 39) = .161$, $p = .691$, $\eta^2 = .004$, 90% CI[0, .037]), feedback valence x time interaction ($F(1, 39) = .433$, $p = .514$, $\eta^2 = .011$, 90% CI[0, .064]), or feedback self-congruence x feedback valence x time interaction ($F(1, 39) = .676$, $p = .416$, $\eta^2 = .017$, 90% CI[0, .083]). Data was also analysed by means of LMMs. The results obtained were consistent with those of the primary analysis.

Discussion

In Experiment 3, we addressed methodological limitations of Experiment 2 to provide a more reliable testing ground for our hypotheses. Our methodological adjustments successfully

eliminated feedback discrepancy asymmetries across the different feedback conditions. Our findings suggested that both our need to gain self-concept positivity and our need to maintain self-concept stability play a role in constraining the integration of self-relevant social feedback.

The primary analyses of Study 3 confirmed the significant effect of feedback self-congruence, mirroring the results of Study 2. This consistency across studies supports our hypothesis that when processing self-relevant feedback, individuals preferentially integrate information that reinforce their current self-concept. Conversely, social inputs that conflict with our self-views seemingly fail to stimulate a feedback-contingent shift in our self-representations, regardless of their valence. Contrary to the non-significant effect of feedback valence in Experiment 2, a moderate effect was observed in Experiment 3. We found that positive feedback was moderately more integrated than negative feedback. A closer examination of the results suggested that not only feedback valence had a lesser impact on participants self-representations than feedback self-congruence, but also greater variability among individuals. Interestingly, the examination of specific update scores under each condition suggested that only self-congruent positive and negative feedback truly stimulated contingent changes in participants self-representations. In contrast, self-incongruent positive feedback seemed to have no effect, while self-incongruent negative feedback involved substantial and directionally inconsistent updates of participants self-representations. Importantly, these updating patterns were found to be stable even 1 day later, suggesting that the observed self-representation changes might represent enduring adjustments to one's self-concept.

Experiment 4

Introduction

Experiment 3 suggested that existing self-concept content plays a critical role in the incorporation of new self-relevant information. Results revealed that individuals tend to preferentially integrate information that aligns with their pre-existing self-concept, and that this effect is more pronounced when the information received has the potential to improve self-

concept positivity. Although these results reconcile previous conflicting findings (Alicke & Sedikides, 2009; Hepper et al., 2011; Swann Jr. & Brooks, 2012; Swann, Tafarodi, et al., 1992), they do not provide any insights into whether the observed effects are specific to self-concept or reflect a more general tendency to form and update beliefs about personality traits, including those of other people. This comparison (self vs. others) has been used as a strategy in a wide range of studies to provide insights into the functioning of our cognition, behavior and affect, and represent a solid approach to test the influence of our self-concept on many cognitive and perceptual domains (Chakraborty & Chakrabarti, 2018; D'Argembeau et al., 2008; Frings & Wentura, 2014; Jenkins & Mitchell, 2011; Ma & Han, 2010; Miles et al., 2010). In turn, the study of differences and similarities in the potential of congruent and incongruent feedback to generate changes in the representations we have about ourselves and others may be critical for understanding the formation and stabilisation of our beliefs, and help clarifying the boundaries between self-related and social cognition.

In our daily experience we constantly encounter new people and interact with them. To effectively navigate our social environments, we should form quick and accurate representations of others' personalities, which involves integrating information about their attributes during the belief formation process (Frolichs et al., 2022). There is also evidence of an asymmetry in favor of positive feedback when integrating information about the traits of others (Korn et al., 2012, 2014, 2016). However, it has been suggested that the mechanisms for updating beliefs about oneself and others are not identical. Although both self- and other-focused belief updating may show a similar preference for positive feedback, the cognitive processes driving these updates might vary depending on whether the focus is on the self or someone else (Korn et al., 2012, 2014).

Critically, differences and similarities in belief updating patterns between the self and others may also emerge as a result of various combinations of the integration of congruent and incongruent feedback. When facing self-related social feedback, the social inputs susceptible

to be integrated into the self have to find its place in a well-grounded self-knowledge base, which contains autobiographical evidence supporting our self-representations (Conway, 2005; Haslam et al., 2011). According to our results, any information that challenges our self-views might be rejected or ignored. In contrast, forming beliefs about others' personalities involves making judgments about their traits based on limited behavioral samples, which may entail a more dynamic process in which initially incongruent information is integrated to form a more precise (though possibly biased) view of others' attributes (Frolichs et al., 2022).

The purpose of this study was to determine the specificity of the effects observed in Experiment 3. We hypothesized that individuals would tend to incorporate more positive than negative feedback from into their beliefs when evaluating the attributes of others. We also anticipated observing a diminished congruence effect, suggesting that individuals may exhibit a decreased resistance to modifying their pre-existing beliefs when updating their perceptions about other people.

Methods

Participants

For this study, we recruited 44 participants. Two participants were excluded from the sample based on their BDI-II scores (score participant 1: 35, score participant 2: 28). One participant was excluded from the sample based on their number of missing responses [>15% (% missing responses: 75%)]. As in study two and three, we also excluded from the sample those participants that did not make a minimum number of negative decisions in the sentence verification task. Two participants did not reach the minimum number of negative decisions required and were excluded from further analysis. The final sample was composed of 40 participants (28 female M_{age}= 21.14, SD_{age}= 1.77).

Procedure

In this study, we adapted the procedure previously used in Experiment 3. Participants listened to recordings of personality descriptions allegedly coming from former participants in the experiment. These recordings, which were approximately 7 minutes in length (mean = 6.87, standard deviation = .44), were made by 7 external collaborators (4 females, 3 males) who were unaware of the purpose of the study in order to maintain authenticity in their descriptions. Recordings were randomized across participants. After listening to the description, participants had to judge how similar they perceived their personality to be to the personality of the person of the recording using a scale from 1 (completely different) to 100 (identical). Participants then completed the sentence verification task, in which they provided dichotomous responses regarding the fit of 95 traits with the personality of the person described in the recording. Using non-proportional stratified random sampling, the program obtained a balanced set of positive and negative categorical decisions. Finally, participants completed the belief updating task, which followed the same structure as in Experiments 2 and 3, with one variation: they were told they would receive feedback from three individuals who had also rated the person in the recording. The sentence preceding feedback in Experiments 1, 2, and 3 "Others think you are: ") was replaced with "Others think this person is: ", followed by each filtered adjective and the corresponding feedback score.

Results

As in study 3, we first checked whether feedback discrepancies did not significantly differ across feedback conditions. We conducted a rmANOVA with average feedback discrepancy as the dependent variable, and Feedback congruence and Feedback valence as within-participants factors. Results showed no significant feedback discrepancy asymmetries across feedback conditions (Feedback self-congruence $F(1, 39) = 1.327$, $p = .256$, $\eta^2 = .032$, 90% CI[0, .120], Feedback valence $F(1, 39) = 1.563$, $p = .219$, $\eta^2 = .038$, 90% CI[0, .133], Feedback Congruence X Feedback valence $F(1, 39) = .328$, $p = .571$, $\eta^2 = .008$, 90% CI[0, .054].

Next, we reproduced the analysis conducted in Experiments 2 and 3. We conducted a rmANOVA with average updating scores as the dependent variable, feedback valence, feedback congruence and their two-way interaction as within-participants factors, and feedback discrepancy and update space as covariates. Results showed a significant and moderate main effect of feedback self-congruence ($F(1, 37) = 5.964$ $p = .019$, $\eta^2 = .138$, 90% CI[.012, .296]), a non-significant main effect of feedback valence ($F(1, 37) = .201$, $p = .657$, $\eta^2 = .005$, 90% CI[.0, .042]) and a significant and strong interaction between feedback self-congruence and feedback valence ($F(1, 37) = 12.994$, $p < .001$, $\eta^2 = .259$, 90% CI[.080, 420]). Next, we followed the same analytical approach and conducted a similar analysis through LMMs. After model selection (BIC) the best model (Maringal $R^2 = .121$, Conditional $R^2 = .205$) included main effects for feedback congruence, feedback valence, feedback discrepancy and ipdate space, an interaction between feedback self-congruence and feedback valence, random intercepts for participants and random slopes for both feedback self-congruence and feedback valence.

To examine the interaction between feedback self-congruence and feedback valence, we conducted pairwise comparisons with Tukey-corrected p-values and Sidak-corrected confidence intervals. Our results showed that there were no significant differences in update scores among the congruent + positive feedback, congruent + negative feedback, and incongruent + positive feedback conditions (all p-values > .8). However, when receiving incongruent + negative feedback, participants updated their representations of others' traits significantly less compared to all other conditions (all $p < .01$). Sidak-corrected CIs showed that participants integrated social evaluative feedback in their representations of others' traits (adjusted CIs did not include 0) when receiving both positive and negative self-congruent feedback (positive feedback: 95% CI[2.047, 7.932], negative feedback: 95% CI[2.263, 8.688]) and when receiving incongruent and positive feedback (95% CI[1.239, 6.967]). However, incongruent and negative feedback did not alter participants beliefs (95% CI[-4.592, 1.037]).

In contrast to Experiment 3, the results of Experiment 4 suggested that changes in participants representations of others' traits where not fully restricted by the congruence of social evaluative feedback, as reflected by participants' integration of incongruent and positive information into their prior beliefs. Moreover, in Experiment 3 participants' bias towards positive feedback was reflected by the preferential and significant integration of positive (vs negative) congruent feedback, the dismissal of incongruent and positive feedback and a substantial rejection of incongruent and negative feedback. However, in this study, the bias towards positive feedback only emerged in the incongruent feedback conditions, where negative feedback was dismissed, and positive feedback was substantially integrated. To provide a direct comparison of results obtained in studies 3 and 4 we gathered the data and created a dichotomous variable representing the difference in the target under judgment (Experiment 3: Self, Experiment 4: Others).

We built a LMM that tried to capture the differences in the effects found between both studies. This model included main effects for feedback congruence, feedback valence, target (Self, Others), feedback discrepancy, update space, and a three-way interaction between feedback congruence, feedback valence, and target. Following the marginality principle, two-way interactions were also included. To see if any alternative model could account better for the data, we compared the fitted model against LMMs including different combinations of random intercepts for participants and traits and random slopes for feedback congruence and feedback valence. BIC determined that the best model (Maringal R^2 = .144, Conditional R^2 = .219) was that including the three-way interaction, a random intercept for participants and random slopes for feedback congruence and feedback valence. A global test on model predictors indicated that the three-way interaction was statistically significant ($F(1,3041.5) = 6.978$, $p = .008$). As evidenced by the significant interaction, participants from study 3 and 4 differed in their updating patterns. To further investigate these differences, we conducted a post-hoc analysis comparing both groups (target) across each feedback condition (Tukey-corrected p-values). Results revealed no significant differences in belief updating between

groups under congruent and negative feedback (t(110) = 1.508, p = .134) or congruent and positive feedback (t(110) = -1.233, p = .221). However, we found that participants in the ''Others'' group incorporated significantly more social evaluative feedback into their prior representations under both incongruent and positive feedback (t(110) = 3.509, p < .001) and incongruent and negative feedback (t(110) = 2.804, p = .005).

Discussion

In Experiment 4, we investigated whether the observed effects when updating beliefs about the self (Experiment 3) would extend to updating beliefs about the personality of others. In line with our hypothesis, our findings indicated that when forming beliefs about others' attributes we display a more flexible approach that allows conflicting information to alter our initial representations. Notably, we also found that participants tended to incorporate more positive than negative feedback about others, in line with previous research (Korn et al., 2012, 2014). However, this effect was only present under incongruent feedback. By examining condition-specific update scores we found that all feedback types, except for incongruent and negative social inputs, elicited feedback-consistent shifts in participants' prior beliefs. Interestingly, while incongruent and negative feedback led to substantial and directionally inconsistent updates in participants beliefs about the self, this same type of social inputs resulted in negligible updates in participants' beliefs about others. These novel findings shed light on potential self-specific constraints on belief formation and change and suggest that we display belief updating patterns especially suited to maintain a positive and stable self-concept.

The results of experiment 4 also provide insights into how we form beliefs about other people's personalities. Prior research has shown that when facing new information, we also display a tendency to incorporate more positive than negative feedback about others. To our knowledge, this study is the first to experimentally control the initial positive bias in both our self-concept and our beliefs about others' personalities and test the effect of feedback congruence. Importantly, our findings highlight that although a positivity bias is evident when

updating representations of both the self and others, the specific patterns of integration of congruent and incongruent information differ between the two. These findings contribute to a deeper understanding of belief updating processes and emphasize the distinct nature of self-beliefs compared to beliefs about others' personalities.

General discussion

In this research, we tested the idea that when updating our self-representations, a balance exists between the need for adaptability in our self-concept to foster positivity and the need for consistency to preserve our existing self-knowledge. We conducted a comprehensive series of experiments focusing on the integration of new self-relevant information. Our aim was to orthogonalize and test the influences of both feedback self-congruence and feedback valence on the updating of self-representations. Our findings revealed that both the need to preserve existing self-knowledge and the desire for a positive self-image impose self-specific and enduring constraints on the integration of new self-relevant information. By considering the simultaneous roles of positivity and stability motives, we gain a deeper understanding of how our self-concept is shaped and maintained over time.

Our study suggests that when facing new self-relevant social feedback, we tend to preferentially integrate information that aligns with our pre-existing self-concept. This is consistent with research suggesting that we strive for social verification of our self-representations (Swann Jr. & Brooks, 2012; Swann Jr. & Buhrmester, 2012). Importantly, our findings go beyond existing evidence suggesting that we display a preference for receiving self-congruent feedback, and seeking like-minded interaction partners, and suggest that we actively use self-congruent information to strengthen our self-representations, potentially maximizing self-concept stability. Our results also align with the notion that self-representations are embedded within a robust system of self-knowledge that helps us selectively process and assimilate new self-congruent information (Conway, 2005; Conway et al., 2004; Haslam et al., 2011; Klein, 2010). Displaying asymmetric integration of self-

congruent vs. self-incongruent feedback might progressively help individuals distinguish between self-descriptive and non-self-descriptive attributes, and improve self-concept clarity (Campbell, 1990). This would have important implications for everyday functioning, as having a clear view of our own attributes helps us generate predictions about our future, plan our actions, select appropriate interaction partners, and preserve our psychological well-being (Becht et al., 2017; Campbell, 1990; Campbell et al., 2003; Lewandowski & Nardone, 2012).

In line with prior research, our results suggested that when facing self-relevant social feedback, we also tend to integrate more favorable than unfavorable information into our self-concept (Korn et al., 2012). This effect draws on the notion that we are strongly motivated to pursue positive self-representations, even at the expense of accuracy (Alicke & Sedikides, 2009; Sedikides & Alicke, 2019; Taylor et al., 1988). However, the inclusion and experimental orthogonalization of the effect of feedback self-congruence in our study provides novel insights that challenge and refine existing research. Although both positivity and stability motives showed self-specific and long-lasting effects on participants' integration of social feedback, our results suggested that the need to preserve existing self-concept might exert a stronger influence. Specifically, we observed larger effect sizes and reduced variability across participants in the effect of feedback self-congruence compared to feedback valence. Our findings challenge the prevailing notion that the pursuit of positive self-representations is a universal and pervasive motivation among psychologically healthy adults (Alicke & Sedikides, 2009) and emphasize the importance of maintaining one's existing self-concept. One possibility is that self-concept stability not only provides robustness to our self-views but is also necessary to add new information to our existing self-knowledge structures (Emery et al., 2015). If the incorporation of positive information were prioritized over self-concept stability, the incorporated information might be as easily dismissed as it was initially accepted, leading to inconsistent self-representations. We suggest that the incorporation of any self-incongruent (positive or negative) changes into the self-concept might necessitate the presence of stability mechanisms (Conway, 2005; Conway et al., 2004). These mechanisms might serve to

preserve the structural integrity of the self-concept during change processes and stabilize the newly acquired self-knowledge.

Although participants prioritized both self-congruent and positive feedback inputs over their counterparts (self-incongruent and negative feedback), our analysis of condition-specific update scores revealed that only self-congruent (positive and negative) evaluations induced feedback-consistent updates in participants' self-representations. The results from the current research revealed a positivity bias within the self-congruent feedback conditions, wherein participants exhibited a stronger inclination to modify their self-representations in response to positive self-congruent feedback compared to negative self-congruent feedback. The same effect was found when examining update scores under self-incongruent feedback levels. However, our results suggested that this positivity bias was driven by a lack of change under self-incongruent and positive feedback and directionally inconsistent updates under self-incongruent and negative feedback.

The specific updating patterns observed in our study offer novel insights that contribute to our understanding of the dynamics of stability and change in the self-concept. To our knowledge, the lack of integration of positive self-incongruent feedback about one's attributes has never been described. We propose that this phenomenon may be driven by the need to avoid self-concept destabilization and maintain accurate predictions about our behaviors and affect (Conway, 2005; Nowak et al., 2000; Steele, 1988). Moreover, the feedback-inconsistent updates observed in response to self-incongruent and negative feedback may further reflect self-concept protective mechanisms (Alicke & Sedikides, 2009; Conway, 2005; Nowak et al., 2000; Swann Jr. & Brooks, 2012). Although self-protection strategies have been widely described, previous studies have not differentiated the effects of feedback self-congruence and feedback valence. This lack of distinction has resulted in overlapping predictions, suggesting that both negative feedback (whether congruent or incongruent) and incongruent feedback (whether positive or negative) would elicit compensatory reactions to protect our self-views. Our findings indicate that these compensatory reactions may be specifically

triggered by feedback that is both self-incongruent and negative. We propose that the integration of negative social inputs that challenge our current self-concept would impose a double penalty. First, it would involve destabilizing well-grounded positive self-representations or compel the integration of a new negative self-representation. Both of these options have the potential to disrupt self-concept stability (Conway, 2005; Nowak et al., 2000). Besides, it would also lead to a reduction in the overall positivity of our self-concept. In this context, our findings are consistent with the extensively described self-protection strategies, but they provide further understanding of the specific circumstances under which these strategies might be triggered.

An important aspect of this research lies in the methodological refinement employed to orthogonalize the effect of stability and positivity motives on the integration of new self-relevant information. This approach aligns with recent empirical and theoretical research that advocates for simultaneously considering both self-concept motives in the study of self-relevant feedback processing (Elder et al., 2022; Mokady & Reggev, 2022). To the best of our knowledge, this is the first study to experimentally isolate these effects, which provides new strategies that might help refine existing findings. One potential example is the study from Korn et al., (2016) which investigated social feedback processing in patients with Borderline Personality Disorder (BPD). The researchers observed a reduced positivity bias in BPD patients compared to healthy controls, with the former showing a greater propensity to adjust their beliefs in response to negative feedback. Although this result is intuitively appealing, it could potentially be attributed to the tendency of BPD patients to maintain more negative self-representations compared to healthy controls (Van Schie et al., 2020). This might lead them to perceive negative feedback as more consistent with their self-concept. Moreover, BPD patients are not only characterized by more negative self-views but also by a greater instability in their self-concept (Kaufman & Meddaoui, 2021). The methodological and analytical strategies we have developed could provide a valuable tool to quantify the influence of various aspects of their self-concept on their reactions to social feedback, which might lead to the

refinement of existing psychotherapeutic strategies. In addition, the approaches employed in this research might be useful to deepen our understanding of self-relevant feedback processing, refining existing findings from behavioral and neuroimaging studies (e.g., Elder et al., 2022; Koban et al., 2023; Korn et al., 2012, 2014; Vanderhasselt et al., 2015; Yang et al., 2016).

Finally, our findings bridge the gap between two opposing perspectives on the primary motivations shaping our self-concept (Kwang & Swann, 2010; Sedikides & Alicke, 2019). These seemingly opposing views have made similar claims on the importance of self-concept stability and self-concept positivity for our psychological well-being. The current results not only suggest that both stability and positivity are important drivers in the development of our self-concept, but they may also shed light on how the two are interconnected (Campbell, 1990). By selectively integrating self-congruent information, we might gain certainty and clarity about our own attributes, and the preferential integration of positive, self-congruent information could serve as a mechanism for enhancing self-concept positivity. We propose that individual differences in these feedback-processing tendencies may accumulate over time, giving rise to distinct levels of self-concept stability and self-concept positivity in the population. This might shed light on potential mechanisms that underlie the relationship between important structural and affective components of the self (Campbell, 1990; DeMarree & Rios, 2014; Wong et al., 2016).

Conclusion

Understanding self-concept dynamics is a complex endeavour that requires an effort to integrate knowledge and methodological approaches from different lines of research in psychology. Here, we aimed to contribute to the field by providing a more nuanced understanding of how our self-concept is shaped and maintained by integrating past effort from social, personality and cognitive psychology. By experimentally disentangling the effects of feedback self-congruence and feedback valence we have shown that when facing new self-

relevant information, there is a trade-off between stabilizing and enhancing our self-views that might provide us with a progressively stable and positive self-concept. We developed methodological and analytical strategies that can be applied to a wide range of studies in the domain of self-concept and social feedback processing. This approach has the potential to refine our understanding of self-concept dynamics and provide new tools for future research.

Acknowledgements

This work was supported by the Spanish Ministerio de Ciencia, Innovación y Universidades, which is part of Agencia Estatal de Investigación (AEI), through the project PID2019-111199GB-I00 to L.F. (Co-funded by European Regional Development Fund. ERDF, a way to build Europe). We thank CERCA Programme/Generalitat de Catalunya for institutional support. We thank Christoph W. Korn and William Swann for their inspiring and helpful feedback.

References

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*(1), 1–48. https://doi.org/10.1080/10463280802613866

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

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. https://doi.org/https://doi.org/10.1016/j.jml.2007.12.005

Baranski, E., Gardiner, G., Lee, D., Funder, D. C., Beramendi, M., Bastian, B., Neubauer, A., Cortez, D., Roth, E., Torres, A., Zanini, D. S., Petkova, K., Tracy, J., Amiot, C. E., Pelletier-Dumas, M., González, R., Rosenbluth, A., Salgado, S., Guan, Y., … Bui, H. T. T. (2021). Who in the World Is Trying to Change Their Personality Traits? Volitional Personality Change Among College Students in Six Continents. *Journal of Personality and Social Psychology*, *121*(5), 1140–1156. https://doi.org/10.1037/pspp0000389

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

Becht, A. I., Nelemans, S. A., van Dijk, M. P. A., Branje, S. J. T., Van Lier, P. A. C., Denissen, J. J. A., & Meeus, W. H. J. (2017). Clear Self, Better Relationships: Adolescents' Self-Concept Clarity and Relationship Quality With Parents and Peers Across 5 Years. *Child Development*, *88*(6), 1823–1833. https://doi.org/10.1111/cdev.12921

Beck, A. T., Steer, R. A., Epstein, N., & Brown, G. (1990). Beck Self-Concept Test. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*(2), 191–197. https://doi.org/10.1037/1040-3590.2.2.191

Boseovski, J. J. (2010). Evidence for "Rose-colored glasses": An examination of the positivity bias in young children's personality judgments. *Child Development Perspectives*, *4*(3), 212–218. https://doi.org/10.1111/j.1750-8606.2010.00149.x

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

Chakraborty, A., & Chakrabarti, B. (2018). Looking at my own face: Visual processing strategies in self-other face recognition. *Frontiers in Psychology*, *9*(FEB). https://doi.org/10.3389/fpsyg.2018.00121

Clare J. Rathbone, C. J. A. M., & Conway, M. A. (2009). Autobiographical memory and amnesia: Using conceptual knowledge to ground the self. *Neurocase*, *15*(5), 405–418. https://doi.org/10.1080/13554790902849164

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

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, *107*(2), 261–288. https://doi.org/10.1037/0033-295X.107.2.261

Conway, M. A., Singer, J. A., & Tagini, A. (2004). The Self and Autobiographical Memory: Correspondence and Coherence. *Social Cognition*, *22*(5), 491–529. https://doi.org/10.1521/soco.22.5.491.50768

Crone, E. A., Green, K. H., Van De Groep, I. H., & Van Der Cruijsen, R. (2022). *A Neurocognitive Model of Self-Concept Development in Adolescence*. https://doi.org/10.1146/annurev-devpsych-120920

D'Argembeau, A., Feyers, D., Majerus, S., Collette, F., Van der Linden, M., Maquet, P., & Salmon, E. (2008). Self-reflection across time: Cortical midline structures differentiate between present and past selves. *Social Cognitive and Affective Neuroscience*, *3*(3), 244–252. https://doi.org/10.1093/scan/nsn020

DeMarree, K. G., & Rios, K. (2014). Understanding the relationship between self-esteem and self-clarity: The role of desired self-esteem. *Journal of Experimental Social Psychology*, *50*(1), 202–209. https://doi.org/10.1016/j.jesp.2013.10.003

Dumas, J. E., Johnson, M., & Lynch, A. M. (2002). Likableness, familiarity, and frequency of 844 person-descriptive words. *Personality and Individual Differences*, *32*(3), 523–531. https://doi.org/https://doi.org/10.1016/S0191-8869(01)00054-X

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why People Fail to Recognize Their Own Incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235

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

Emery, L. F., Walsh, C., & Slotter, E. B. (2015). Knowing Who You Are and Adding to It: Reduced Self-Concept Clarity Predicts Reduced Self-Expansion. *Social Psychological and Personality Science*, *6*(3), 259–266. https://doi.org/10.1177/1948550614555029

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

Frings, C., & Wentura, D. (2014). Self-priorization processes in action and perception. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 1737–1740. https://doi.org/10.1037/a0037376

Frolichs, K. M. M., Rosenblau, G., & Korn, C. W. (2022). Incorporating social knowledge structures into computational models. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-33418-2

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

Grilli, M. D. (2017). The association of personal semantic memory to identity representations: insight into higher-order networks of autobiographical contents. *Memory*, *25*(10), 1435–1443. https://doi.org/10.1080/09658211.2017.1315137

Grilli, M. D., & Verfaellie, M. (2014). Supporting the self-concept with memory: Insight from amnesia. *Social Cognitive and Affective Neuroscience*, *10*(12), 1684–1692. https://doi.org/10.1093/scan/nsv056

Haslam, C., Jetten, J., Haslam, S. A., Pugliese, C., & Tonks, J. (2011). “I remember therefore I am, and I am therefore I remember”: Exploring the contributions of episodic and semantic self-knowledge to strength of identity. *British Journal of Psychology*, *102*(2), 184–203. https://doi.org/10.1348/000712610X508091

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

Hepper, E. G., Hart, C. M., Gregg, A. P., & Sedikides, C. (2011). Motivated expectations of positive feedback in social interactions. *Journal of Social Psychology*, *151*(4), 455–477. https://doi.org/10.1080/00224545.2010.503722

Higgins, E. T. (1987). Self-Discrepancy: A Theory Relating Self and Affect. *Psychological Review*, *94*(3), 319–340. https://doi.org/10.1037/0033-295X.94.3.319

Jenkins, A. C., & Mitchell, J. P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social Neuroscience*, *6*(3), 211–218. https://doi.org/10.1080/17470919.2010.507948

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

Kim, Y. H., & Chiu, C. Y. (2011). Emotional costs of inaccurate self-assessments: Both self-effacement and self-enhancement can lead to dejection. *Emotion*, *11*(5), 1096–1104. https://doi.org/10.1037/a0025478

Klein, S. B. (2010). The self: As a construct in psychology and neuropsychological evidence for its multiplicity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(2), 172–183. https://doi.org/10.1002/wcs.25

Koban, L., Andrews-Hanna, J. R., Ives, L., Wager, T. D., & Arch, J. J. (2023). Brain mediators of biased social learning of self-perception in social anxiety disorder. *Translational Psychiatry*, *13*(1). https://doi.org/10.1038/s41398-023-02587-z

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

Kwang, T., & Swann, W. B. (2010). Do People Embrace Praise Even When They Feel Unworthy? A Review of Critical Tests of Self-Enhancement Versus Self-Verification. *Personality and Social Psychology Review*, *14*(3), 263–280. https://doi.org/10.1177/1088868310365876

Lewandowski, G. W., & Nardone, N. (2012). Self-concept Clarity's Role in Self-Other Agreement and the Accuracy of Behavioral Prediction. *Self and Identity*, *11*(1), 71–89. https://doi.org/10.1080/15298868.2010.512133

Lenth R. Emmeans: estimated marginal means, aka least-squares means. R package version 1.4.7. 2020. Available at: https://CRAN.R-project.org/package=emmeans.

Libby, L. K., & Eibach, R. P. (2002). Looking back in time: Self-concept change affects visual perspective in autobiographical memory. In *Journal of Personality and Social Psychology* (Vol. 82, Issue 2, pp. 167–179). American Psychological Association Inc. https://doi.org/10.1037/0022-3514.82.2.167

Ma, Y., & Han, S. (2010). Why we respond faster to the self than to others? An implicit positive association theory of self-advantage during implicit face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(3), 619–633. https://doi.org/10.1037/a0015797

Manzi, C., Vignoles, V. L., & Regalia, C. (2010). Accommodating a new identity: Possible selves, identity change and well-being across two life-transitions. *European Journal of Social Psychology*, *40*(6), 970–984. https://doi.org/10.1002/ejsp.669

Markus, H., & Wurf, E. (1986). *THE DYNAMIC SELF-CONCEPT: A Social Psychological Perspective*. www.annualreviews.org

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. In *British Journal of Educational Psychology* (Vol. 81, Issue 1, pp. 59–77). https://doi.org/10.1348/000709910X503501

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

Miles, L. K., Nind, L. K., Henderson, Z., & Macrae, C. N. (2010). Moving memories: Behavioral synchrony and memory for self and others. *Journal of Experimental Social Psychology*, *46*(2), 457–460. https://doi.org/https://doi.org/10.1016/j.jesp.2009.12.006

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

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).

Preuss, G. S., & Alicke, M. D. (2009). Everybody Loves Me: Self-Evaluations and Metaperceptions of Dating Popularity. *Personality and Social Psychology Bulletin*, *35*(7), 937–950. https://doi.org/10.1177/0146167209335298

Pyszczynski, T., Greenberg, J., & LaPrelle, J. (1985). Social comparison after success and failure: Biased search for information consistent with a self-serving conclusion. *Journal of Experimental Social Psychology*, *21*(2), 195–211. https://doi.org/https://doi.org/10.1016/0022-1031(85)90015-0

Rathbone, C. J., & Moulin, C. J. A. (2014). Measuring Autobiographical Fluency in the Self-Memory System. *Quarterly Journal of Experimental Psychology*, *67*(9), 1661–1667. https://doi.org/10.1080/17470218.2014.913069

Rathbone, C. J., Moulin, C. J. A., & Conway, M. A. (2008). Self-centered memories: The reminiscence bump and the self. *Memory and Cognition*, *36*(8), 1403–1414. https://doi.org/10.3758/MC.36.8.1403

Reitz, A. K., Zimmermann, J., Hutteman, R., Specht, J., & Neyer, F. J. (2014). How Peers Make a Difference: The Role of Peer Groups and Peer Relationships in Personality Development. *European Journal of Personality*, *28*(3), 279–288. https://doi.org/10.1002/per.1965

Renoult, L., Davidson, P. S. R., Palombo, D. J., Moscovitch, M., & Levine, B. (2012). Personal semantics: At the crossroads of semantic and episodic memory. In *Trends in Cognitive Sciences* (Vol. 16, Issue 11, pp. 550–558). https://doi.org/10.1016/j.tics.2012.09.003

Robinson, D. T., & Smith-Lovin, L. (1992). Selective Interaction as a Strategy for Identity Maintenance: An Affect Control Model. *Social Psychology Quarterly*, *55*(1), 12–28. https://doi.org/10.2307/2786683

Rodman, A. M., Powers, K. E., & Somerville, L. H. (2017). Development of self-protective biases in response to social evaluative feedback. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(50), 13158–13163. https://doi.org/10.1073/pnas.1712398114

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. http://www.jstor.org/stable/2958889

Sedikides, C., & Alicke, M. D. (2019). The five pillars of self-enhancement and self-protection. In *The Oxford handbook of human motivation, 2nd ed.* (pp. 307–319). Oxford University Press.

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

Steele, C. M. (1988). The Psychology of Self-Affirmation: Sustaining the Integrity of the Self. *Advances in Experimental Social Psychology*, *21*(C), 261–302. https://doi.org/10.1016/S0065-2601(08)60229-4

Swann, W. B., Jr. (2012). Self-verification theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), Handbook of theories of social psychology (Vol. 2, pp. 23–42). Thousand Oaks, CA: Sage Ltd. http://dx.doi.org/10.4135/9781446249222.n27

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 Jr., W. B., & Buhrmester, M. D. (2012). Self-verification: The search for coherence. In *Handbook of self and identity, 2nd ed.* (pp. 405–424). The Guilford Press.

Swann, W. B., Ely, R. J., Sullivan, M., Batiste, R., & Woznicki, K. (1984). A Battle of Wills: Self-Verification Versus Behavioral Confirmation. In *Journal of Personality and Social Psychology* (Vol. 46, Issue 6).

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., Pelham, B. W., Krull, D. S., & Swann, L. B. (1989). Agreeable Fancy or Disagreeable Truth? Reconciling Self-Enhancement and Self-Verification. In *Journal of Personality and Social Psychology* (Vol. 57, Issue 5).

Swann, W. B., Stein-Seroussi, A., & Giesler, R. B. (1992). Why People Self-Verify. *Journal of Personality and Social Psychology*, *62*(3), 392–401. https://doi.org/10.1037/0022-3514.62.3.392

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).

van Schie, C. C., Chiu, C. De, Rombouts, S. A. R. B., Heiser, W. J., & Elzinga, B. M. (2018). When compliments do not hit but critiques do: An fMRI study into self-esteem and self-knowledge in processing social feedback. *Social Cognitive and Affective Neuroscience*, *13*(4), 404–417. https://doi.org/10.1093/scan/nsy014

Van Schie, C. C., Chiu, C. De, Rombouts, S. A. R. B., Heiser, W. J., & Elzinga, B. M. (2020). Stuck in a negative me: FMRI study on the role of disturbed self-views in social feedback processing in borderline personality disorder. *Psychological Medicine*, *50*(4), 625–635. https://doi.org/10.1017/S0033291719000448

Vanderhasselt, M. A., Remue, J., Ng, K. K., Mueller, S. C., & De Raedt, R. (2015). The regulation of positive and negative social feedback: A psychophysiological study. *Cognitive, Affective and Behavioral Neuroscience*, *15*(3), 553–563. https://doi.org/10.3758/s13415-015-0345-8

Wong, A. E., Vallacher, R. R., & Nowak, A. (2016). Intrinsic dynamics of state self-esteem: The role of self-concept clarity. *Personality and Individual Differences*, *100*, 167–172. https://doi.org/10.1016/j.paid.2016.05.024

Yang, J., Xu, X., Chen, Y., Shi, Z., & Han, S. (2016). Trait self-esteem and neural activities related to self-evaluation and social feedback. *Scientific Reports*, *6*. https://doi.org/10.1038/srep20274

Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2019). The Better-Than-Average Effect in Comparative Self- Evaluation: A Comprehensive Review and Meta-Analysis. *Psychological Bulletin*. https://doi.org/10.1037/bul0000218