

# Journal of Personality and Social Psychology

## Mapping the Self: A Network Approach for Understanding Psychological and Neural Representations of Self-Concept Structure

Jacob Elder, Bernice Cheung, Tyler Davis, and Brent Hughes

Online First Publication, July 4, 2022. <http://dx.doi.org/10.1037/pspa0000315>

### CITATION

Elder, J., Cheung, B., Davis, T., & Hughes, B. (2022, July 4). Mapping the Self: A Network Approach for Understanding Psychological and Neural Representations of Self-Concept Structure. *Journal of Personality and Social Psychology*. Advance online publication. <http://dx.doi.org/10.1037/pspa0000315>

## INNOVATIONS IN SOCIAL PSYCHOLOGY

# Mapping the Self: A Network Approach for Understanding Psychological and Neural Representations of Self-Concept Structure

Jacob Elder<sup>1</sup>, Bernice Cheung<sup>2</sup>, Tyler Davis<sup>3</sup>, and Brent Hughes<sup>1</sup>

<sup>1</sup> Department of Psychology, University of California, Riverside

<sup>2</sup> Department of Psychology, University of Oregon

<sup>3</sup> Independent Researcher, Newark, California, United States

How people self-reflect and maintain a coherent sense of self is an important question that spans from early philosophy to modern psychology and neuroscience. Research on the self-concept has not yet developed and tested a formal model of how beliefs about dependency relations amongst traits may influence self-concept coherence. We first develop a network-based approach, which suggests that people's beliefs about trait relationships contribute to how the self-concept is structured (Study 1). This model describes how people maintain positivity and coherence in self-evaluations, and how trait interrelations relate to activation in brain regions involved in self-referential processing and concept representation (Study 2 and Study 3). Results reveal that a network-based property theorized to be important for coherence (i.e., outdegree centrality) is associated with more favorable and consistent self-evaluations and decreased ventral medial prefrontal cortex (vmPFC) activation. Further, participants higher in self-esteem and lower in depressive symptoms differentiate between higher and lower centrality positive traits more in self-evaluations, reflecting associations between mental health and how people process perceived trait dependencies during self-reflection. Together, our model and findings join individual differences, brain activation, and behavior to present a computational theory of how beliefs about trait relationships contribute to a coherent, interconnected self-concept.

**Keywords:** self-representation, motivation, medial prefrontal cortex, network analysis, trait semantics


**Supplemental materials:** <https://doi.org/10.1037/pspa0000315.supp>


The nature of self-knowledge and how people achieve structured, coherent self-beliefs is a question that has intrigued thinkers for centuries. While Descartes (1641/1998) postulated that the self is a unitary, indivisible substance to which we have clear and unrestricted access through reflection, Hume (1739/2003) suggested that the self emerges through a system of mental representations. Early theorizing on the self has tended to follow Descartes (1641/1998), describing self-beliefs as monolithic and unitary, with later neuroscientific work characterizing the brain regions involved in general self-referential processing rather than considering the structure and interrelatedness of self-beliefs. Although many social psychological theories

acknowledge that the self-concept is a more intricately interconnected and multifaceted structure, there has yet to be a formal and empirically tested model that explains how the interdependent connections between self-beliefs contribute to self-evaluations and self-concept coherence. Here, we develop a normative trait dependency network to quantify the complexity and structure of the trait knowledge that contributes to self-referential processing.

Inspired, in part, by philosophers like Descartes (1641/1998) and Hume (1739/2003), the self-concept has long been a focus of social psychological research. The self-concept is considered a dynamic structure that organizes self-relevant knowledge and experience,

Jacob Elder  <https://orcid.org/0000-0002-5305-7006>

Bernice Cheung  <https://orcid.org/0000-0002-8925-4917>

Brent Hughes  <https://orcid.org/0000-0001-8732-8727>

The authors have no known conflict of interest to disclose. Data, analysis code, and research materials for both the behavioral and the neuroimaging analyses are available on the open science framework at [https://osf.io/3pcvt/?view\\_only=28df28d593354c5a9d65194441877a89](https://osf.io/3pcvt/?view_only=28df28d593354c5a9d65194441877a89). The neuroimaging statistical maps are located at <https://neurovault.org/collections/CRTMNQFW/>. The hypotheses were preregistered for Study 3 but not Study 2, and the preregistration is located at <http://aspredicted.org/blind.php?x=89ag6d>.

Jacob Elder played lead role in data curation, formal analysis, validation, visualization, and writing of original draft; supporting role in

conceptualization; and equal role in methodology, project administration, and writing of review and editing. Bernice Cheung played supporting role in data curation, formal analysis, methodology, validation, visualization, and writing of review and editing and equal role in investigation and project administration. Tyler Davis played supporting role in formal analysis, supervision, and writing of original draft and equal role in conceptualization, methodology, and writing of review and editing. Brent Hughes played lead role in funding acquisition and supervision; supporting role in formal analysis and writing of original draft; and equal role in conceptualization, methodology, and writing of review and editing.

Correspondence concerning this article should be addressed to Brent Hughes, Department of Psychology, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, United States. Email: [bhughes@ucr.edu](mailto:bhughes@ucr.edu)

and guides individuals through inter and intrapersonal processes (Leary & Tangney, 2003; Markus & Wurf, 1987; Vazire & Wilson, 2012). Research suggests that people maintain generally positive and favorable self-views (Sedikides & Gregg, 2008; Taylor & Brown, 1988) and achieve a coherent and stable self-concept by integrating various self-aspects and aligning experiences with self-views (Greenwald, 1980; Swann et al., 2003). Prior research has made analogies to formal computational models of knowledge representation when considering self-concepts, such as associative semantic networks (Anderson & Bower, 1973; Bower & Gilligan, 1979) or hierarchical knowledge structures (Kihlstrom et al., 1988; Marsh & Shavelson, 1985; McConnell, 2011; Schell et al., 1996). Information represented in the self-concept ranges from idiosyncratic, personal memories, and beliefs about oneself that differentiate us from others, to general semantic information that describes how traits relate to each other (Kihlstrom et al., 2003). Here, we develop a formal model of the self-concept that focuses on how people's normative beliefs about trait relationships may help to explain how self-concept positivity and coherence emerge from this structure of trait beliefs, and how this structure shapes self-perceptions.

Formalizing how normative beliefs about directed trait relationships contribute to self-evaluations may help to develop stronger connections between psychological theories on the self-concept and how self-referential processing is reflected in brain function. Past research suggests that cortical midline structures are consistently involved in self-referential processing (Araujo et al., 2013; Denny et al., 2012; Northoff et al., 2006; Wagner et al., 2012). For example, the medial prefrontal cortex (mPFC) is recruited by tasks that require people to reflect on their own personality traits (Craig et al., 1999; Jenkins & Mitchell, 2011; Johnson et al., 2002; Kelley et al., 2002; Moran et al., 2006; Rameson et al., 2010). The ventral mPFC (vmPFC), in particular, may be tuned to features of trait knowledge (Ma et al., 2014) such as positivity (Chavez et al., 2017; Hughes & Beer, 2013), personal importance (D'Argembeau, 2013), and breadth (Beer & Hughes, 2010). In separate, but related work, the vmPFC is involved in coding for latent causes (Chan et al., 2016), representing task-related cognitive maps (Park et al., 2020; Schuck et al., 2016; Wikenheiser & Schoenbaum, 2016), and encoding and organizing concept information (Constantinescu et al., 2016; Mack et al., 2020). By developing a formal model of self-referential processing that considers people's beliefs about trait dependency relations, it may be possible to bridge these disparate observations from the social and cognitive neuroscience literatures and develop a broader understanding of how the vmPFC contributes to self-concept representation.

In our trait dependency model, traits are connected to each other in a directed network where each connection describes a dependency relation between a pair of traits. These connections were constructed from an independent sample of raters that evaluated, and thus can be viewed as reflecting beliefs about trait dependency relations that people in the study population generally agree upon (Study 1). For example, our independent raters judged the trait "witty" as depending on "fun," and if enough people agreed upon this dependency relation, a connection pointing from "fun" to "witty" was drawn in the network. Likewise, "fun" was judged as depending on "sociable," which in turn depended on "outgoing." By connecting these traits in a directed chain, the model provides insight into the dependency relationships between traits that people feel normatively committed to. For example, if someone describes

themselves as "outgoing," they may also describe themselves as "sociable," "fun," or "witty," as people generally believe that these traits depend upon "outgoing." Furthermore, if someone describes themselves as "witty," they ought to at least partially endorse traits that they believe "witty" depends on, such as "fun," "sociable," or "outgoing," as failing to do so may create a contradiction that threatens the coherence of the self-concept structure. Critically, this is a model of people's *beliefs about* dependency relations amongst traits and not actual causal relationships or associations amongst traits as they exist in the world. That is, our model posits that people achieve self-concept coherence by avoiding contradictions between self-beliefs on traits that they perceive as depending on one another. In this way, our model is akin to a psychological instantiation of the coherence theory of truth in philosophy (Davidson & LePore, 1986), which posits that truth is a function of having noncontradictory beliefs amongst the global set of propositions one endorses.

The idea of a trait dependency network as a way of representing traits within the self-concept follows from related work on concepts where structured dependency networks are one format for conceptual representations (Hadjichristidis et al., 2004; Rehder, 2003; Sloman et al., 1998). In such dependency network models, distinctions are made between central features (i.e., traits) that have many perceived dependencies and more peripheral features that have fewer perceived dependencies. Central features are key to preserving the stability of a network and tend to be rated as less "mutable," as changing them would potentially have broader effects on the stability and coherence of a network (Sloman et al., 1998). For example, for the concept "robin," the feature "wings" might be central, as other features like "flying" and "nest building" might depend on a bird having "wings," whereas the feature "red breast" is less central as the status of these other features do not depend on breast color.

There are a number of different types of centrality, but here we focus primarily on "outdegree" centrality, which simply describes how many other traits are perceived as depending on a given trait. Translated to self-evaluations, higher outdegree traits (i.e., traits perceived as having more dependencies) should be more critical to people's self-concept coherence, as self-evaluations on downstream, dependent traits should remain consistent with the upstream trait that they depend on. Endorsing a downstream trait and not an upstream trait that it depends on could result in a contradiction that disrupts coherence. Likewise, higher outdegree traits should be critical to self-concept positivity, as positive evaluations on such traits permit positive evaluations of their downstream dependents without contradiction, thus propagating positivity across network connections. Importantly, outdegree centrality is a property of traits that reflects the role these traits play in global coherence but is not coherence per se, as coherence is only a property of the whole system of self-beliefs. We compare outdegree centrality to another local centrality measure, "indegree" centrality, which describes how many traits a given trait is perceived as depending on. Unlike higher outdegree centrality traits, higher indegree centrality traits are not, on average, as critical to maintaining coherence, as they are more influenced by other traits rather than exerting an influence on other traits.

By making quantifiable predictions for how traits are connected in the self-concept, our trait dependency network can integrate semantics, brain, and behavior, into a common theoretical framework that informs both social psychological and neuroscience theory. As a first test of this model, we investigate how people's beliefs about

trait dependencies may relate to their tendencies to represent a positive and coherent self-concept. Specifically, in behavioral (Study 2) and neuroimaging (Study 3) studies, we tested the hypothesis that people would self-evaluate more favorably on traits (i.e., more self-descriptive on positive traits and less self-descriptive on negative traits) as a function of dependencies (i.e., outdegree centrality). To bolster this primary hypothesis, we also tested whether people do, empirically, maintain more consistency with higher outdegree traits by examining the extent to which traits are evaluated similarly to their neighbors (i.e., traits that have a direct dependency relationship, and are therefore adjacent in the network). If people are using higher outdegree traits to maintain a stable, coherent self-concept, these traits should be evaluated more consistently with their neighbors than lower outdegree traits.

If more favorable self-evaluations on higher outdegree traits are critical for maintaining positivity and coherence of the self-concept, individual differences in how self-evaluations track outdegree centrality could also be important for mental health and well-being (Studies 2 and 3). The tendency to view oneself positively is important for psychological adjustment, and this relationship is robust across sex, age, and culture (Dufner et al., 2019). Disruptions in this tendency are related to depressive symptoms (Alloy et al., 2012), with depressed individuals seeking negative over positive feedback (Giesler et al., 1996) and maintaining inflexible negative self-views (Malle & Horowitz, 1995; Stange et al., 2017). Beyond the positivity of self-views, other research shows that the structure and organization of self-views relates to psychological adjustment (McConnell & Strain, 2007). Individuals higher in depression and lower in self-esteem have more compartmentalized, rather than integrated, positive and negative self-aspects (Showers, 1992; Showers et al., 2015; Stopa et al., 2010) and fewer and less complex self-aspects (Linville, 1987; McConnell et al., 2005). Therefore, both the positivity and structure of self-views may be important to psychological adjustment and mental health outcomes. In terms of our trait dependency network model, individuals with higher self-esteem and fewer depressive symptoms (Orth et al., 2009; Orth & Robins, 2013; Steiger et al., 2014) may be more attuned to normative trait relationships that help preserve self-concept coherence and positivity, such as a trait's outdegree centrality. Conversely, being indiscriminate toward trait dependencies may be psychologically hazardous, as this may lead to inconsistencies between related self-beliefs and to an inability to effectively maintain positive self-views between dependent traits. Consequently, this may lead to disruptions in coherence and positivity in the broader self-concept structure.

In addition to behavior and individual differences, how brain processing reflects perceived trait dependency relations during self-evaluations may provide insight into the neural mechanisms supporting self-concept organization and inference (Study 3). First, we predicted that vmPFC—a region theorized to be critical for self-evaluations (D'Argembeau, 2013) and general conceptual representation (Mack et al., 2020)—would track the outdegree centrality of traits. Next, to test how the brain represents trait dependency relationships that contribute to self-concept structure, we examined how the pairwise similarity of traits, as determined by the number of their shared neighbors in the network, relates to voxelwise activation patterns. We predicted regions that encode semantic and trait knowledge (e.g., anterior temporal lobe), and secondarily regions that are involved in self-referential processing (e.g., mPFC), would exhibit neural response patterns reflecting the similarity of trait

relationships. If the model captures aspects of how the brain represents the self-concept, then brain activation patterns should reflect the fine-grained similarity relationships between traits in the network, in addition to higher-order measures like centrality. By demonstrating how the trait dependency network relates to brain and behavior, our approach allows us to examine the interconnected nature of the self-concept, and provides insight into how structured trait knowledge shapes self-processing.

## Study 1

The primary goal of Study 1 was to collect dependency ratings for a set of trait adjectives to generate our trait dependency network. As described above, this trait dependency network is a normative model of people's beliefs about trait dependencies, which we use to make predictions for new independent participants in Study 2 and Study 3. After forming the network from the ratings of participants, our network measures like outdegree centrality, indegree centrality, and pairwise similarity can be leveraged to test how the network representation is reflected in behaviors and brain activation. In addition to the primary dependency measures, we also collected additional normative ratings on several trait characteristics (e.g., the extent to which traits are desirable, broad, interpersonal, externally observable, and prevalent) for validation purposes.

## Method

### Participants

One hundred seventy-eight participants (65% male, 35% female) were recruited from Amazon Mechanical Turk from ages 19–75 ( $M_{\text{age}} = 32.76$ ) for the positive network development. The positive network participants were 56.5% Caucasian, 20.3% Asian, 10.2% African-American, 9.0% Hispanic, 3.9% mixed/other. One hundred seventy-eight participants (66% male, 34% female) from ages 19–71 ( $M_{\text{age}} = 32.79$ ) were recruited for the negative network development. The negative network participants were 48.0% Caucasian, 23.7% Asian, 16.4% African-American, 4.5% Hispanic, and 7.3% mixed/other. Two hundred participants (72% male, 28% female) from ages 20–67 ( $M_{\text{age}} = 31.29$ ) were recruited for the normative ratings. The normative rating participants were 53.5% Asian, 31.0% Caucasian, 8% African-American, 4.0% Hispanic, and 2.0% mixed, 1.5% missing.

### Procedure

In order to construct a network of dependency relationships between trait words, we first generated a list of 292 positive traits and 332 negative traits (Anderson, 1968; Hampson et al., 1987; Kirby & Gardner, 1972).

**Normative Ratings and Trait Selection.** Participants rated either all positive traits or all negative traits on one of the following five dimensions: desirability ("Please rate the extent to which it would be desirable for an individual to possess each trait," from 1 *extremely undesirable* to 7 *extremely desirable*), category breadth ("Please rate the extent to which a trait is broad," from 1 *extremely specific* to 7 *extremely broad*), interpersonal ("Please indicate for each trait the extent to which it describes an interpersonal quality, that is the extent to which a trait describes how one person relates or interacts with other people," from 1 *not at all interpersonal* to

7 *extremely interpersonal*), observability (“Please rate the extent to which a trait is observable,” from 1 *extremely difficult to observe* to 7 *extremely easy to observe*), and prevalence (“How frequently would you expect to see this trait in the general population,” from 1 *very infrequently* to 7 *very frequently*). Each normative rating of positive or negative valence was judged by 20 raters. We then ran 100,000 simulations, and in each simulation, we randomly drew 150 traits out of the entire list of possible traits and calculated a two-way average consistency intraclass correlation coefficient (ICC;  $C, k$ ), for each dimension, testing for consistency in normative ratings on the traits. More specifically, given that ICCs measure interrater reliability, the goal of these simulations was to identify a subset of traits from the larger sample of traits that raters provided the most consistent normative ratings for. To do so, different samples of 150 traits were drawn from the larger sample of traits at each iteration, and an ICC was estimated for that sample of traits based on the raters’ consistency in judgments. This was performed for each normative characteristic at each iteration (desirability, observability, interpersonal, prevalence, and breadth). Among the 100,000 different subsets of traits, we identified which of those samples of traits had at least three normative characteristics with an ICC > 0.7. This ensures that there was sufficient consistency among raters on at least three of the five normative characteristics for all traits selected. Subsequently, we chose the list that had the fewest outliers in desirability ratings and the most interpretable traits. Interpretability was based off judgments from three separate reviewers who independently flagged traits that were either considered uncommon or had ambiguous meanings and traits were omitted if 2 out of 3 reviewers flagged a trait. This procedure allowed us to acquire a list of 150 positive traits and 150 negative traits. Two trait words were deleted from positive valence for having a mean desirability below the midpoint 4, and two trait words were deleted from negative valence for having a mean desirability above the midpoint 4, to arrive at a total of 148 positive traits and 148 negative traits (see Supplemental Table 1 for all traits; Supplemental Table 2 for positive traits sorted by outdegree centrality; Supplemental Table 3 for negative traits sorted by outdegree centrality).

**Network Construction.** We collected ratings of dependency amongst the traits from separate groups of participants for positive and negative traits. On each trial, participants were presented with one trait word as a target trait. Participants were then presented with a list of the remaining 147 trait words and asked, “Which traits does [TARGET TRAIT] depend upon?” Participants were able to nominate as many trait words as they believed were applicable. Participants completed a total of 10 trials with 10 randomly selected traits as the target trait. This procedure was executed for all 148 positive traits and 148 negative traits, and at the conclusion, 144 positive words were evaluated 12 times as the target word and four were evaluated 13 times. Adjacency matrices ( $A$ ), describing the dependency relations nominated by participants, were calculated separately for positive and negative traits. The matrices were 148 columns (dependents) by 148 rows (dependencies), with each cell ( $A_{ij}$ ) describing whether the trait in column  $j$  is rated as dependent on the trait in row  $i$  (e.g., if a participant nominates “nice” as dependent on “outgoing,” then a 1 is added to the matrix cell  $A_{\text{Outgoing} \times \text{Nice}}$  at the column for “nice” and row for “outgoing”). To avoid including dependencies for terms that were rarely or spuriously endorsed, we set an *a priori* threshold, based on a

previous pilot study (Davis et al., 2014) requiring at least 25% of participants to endorse a dependency before adding it to the adjacency matrix. Therefore, each entry to the adjacency matrix is 1 or 0 depending on whether the number of participants who judged a given trait to be dependent on another exceeded our 25% threshold. Given that this network is a normative network requiring a certain degree of consensus, it does not necessarily reflect people’s individual beliefs about dependencies and there may be idiosyncratic differences from our network, but it should *on average* reflect people’s individual beliefs about dependency relations. Requiring a moderate amount of consensus on dependencies may also help to reduce the influence of individual differences (e.g., personality, demographic differences) on network structure.

**Stability Across Thresholds.** First, we assessed the reliability of the network predictions across possible choices of threshold to test whether our *a priori* threshold of 25% endorsements was justified. To this end, we varied the threshold from 5% to 75% at increments of 5%, recomputed the adjacency matrix at each threshold, and then recomputed our primary measures. Stability was calculated as the correlation between a measure computed at a given threshold with those computed at all other thresholds.

**Bootstrapped Reliability.** As an additional way of testing the reliability of the network measures across raters, we adopted a bootstrap procedure where we iteratively resampled the network using subsets of raters. A bootstrap procedure was adopted instead of common internal consistency tests for reliability (e.g., Cronbach’s alpha, McDonald’s omega) because the networks measures are not simply linear combinations of trait ratings. We sampled with replacement from each of the 12–13 participants who nominated for each of the 296 traits, and at each resample, reconstructed a new network, and computed outdegree centrality, indegree centrality, and similarity measures. Correlations were computed between the bootstrapped centrality and similarity measures and those calculated from the original network. This procedure was iterated for 20,000 simulations yielding a distribution of correlations for each network measure reflecting how reliable that measure is across resamples.

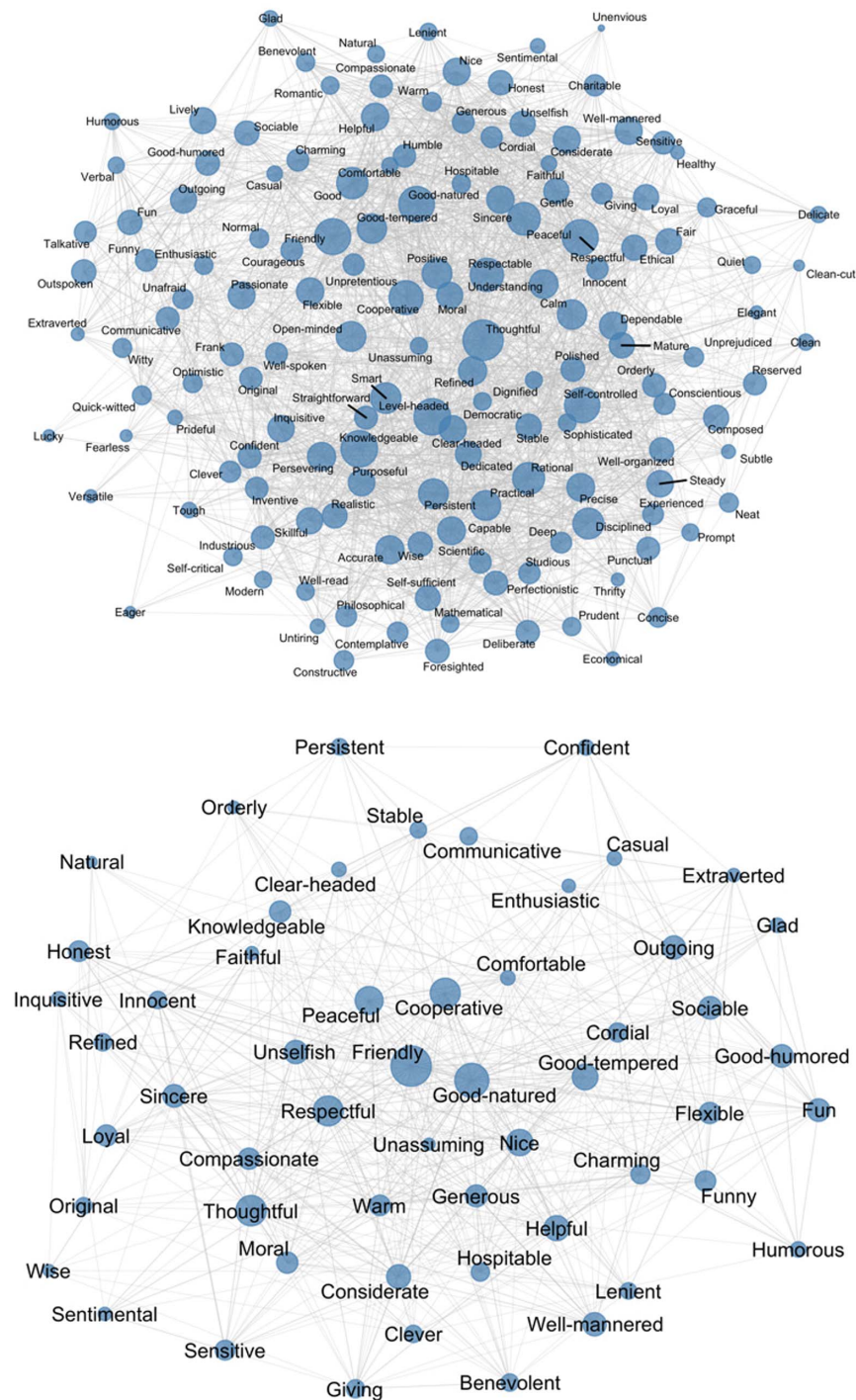
**Network Measures.** We calculated several network statistics from the network adjacency matrices (see Figure 1 for the network). *Outdegree centrality* was defined as the number of traits that depend on a given trait (sum of a given trait’s row in the adjacency matrix; how many of columns  $j$  depend on row  $i$ ). *Indegree centrality* was defined as the number of traits a given trait depends on (sum of a given trait’s column in the adjacency matrix; how many of rows  $i$  column  $j$  depends on). *Pairwise similarity* between traits was calculated as the number of common neighbors between two traits (i.e., how many traits the pair of traits share in being connected to), weighted by the inverse logarithm of their degrees (total number of connections).

### Transparency and Openness

We describe the cleaning and data generating procedures, as well as the network modeling and procedures. We describe how data was collected and provide sufficient information and materials to reproduce results. Data, analysis code, and research materials are available at [https://osf.io/3pcvt/?view\\_only=28df28d593354c5a9d65194441877a89](https://osf.io/3pcvt/?view_only=28df28d593354c5a9d65194441877a89).



**Figure 1**  
*Visual Depiction of the Dependency Trait Network Constructed From Independent Sample of Raters*



*Note.* Participants nominate which traits “respectful” depends on. If over 25% of participants nominated “friendly,” this would result in an output connection from “friendly” to “respectful.” (Top) Full trait network of 148 traits in the positive trait network. Node size varies according to outdegree centrality. (Bottom) Traits dependent on high outdegree centrality trait “friendly.” See Supplemental Table 1 for a table of all traits included in positive and negative networks. See the online article for the color version of this figure.

## Results

### Stability Across Thresholds

Our primary interest in Study 1 was whether the network and the measures obtained from it were stable across different choices of threshold and across raters. Results suggest that our measures are stable across a range of thresholds and support our *a priori* threshold of 25%. The correlations at the 25% threshold are reported for positive outdegree centrality ( $r = .74$ ), negative outdegree centrality ( $r = .65$ ), positive similarity ( $r = .55$ ), and negative similarity ( $r = .51$ ). For a visualization of the network stability results, see Supplemental Figure 1.

### Reliability Across Resamples

Using our bootstrapping procedure to test reliability, we observe very strong reliability for the outdegree centrality of positive,  $r(146) = .94$ , [.92, .95], and negative,  $r(146) = .91$ , [.89, .93], traits, good reliability for the similarity of positive,  $r(21,902) = .73$ , [.62, .8], and negative,  $r(21,902) = .68$ , [.58, .76], traits, and moderate reliability for the indegree centrality of positive,  $r(146) = .54$ , [.39, .68], and negative,  $r(146) = .68$ , [.58, .76], traits. For figures of bootstrapped reliability distributions, see Supplemental Figure 2.

### Associations With Trait Characteristics

To understand how our network measures are related to other possible known normative characteristics of traits, we tested the association between our network measures and the trait dimensions collected in a separate group of participants (desirability, category breadth, interpersonal, observability, and prevalence). First, there was no overall difference between positive and negative traits for outdegree,  $t(294) = .792$ ,  $p = .429$ , or indegree centrality,  $t(294) = .763$ ,  $p = .446$ , suggesting that neither measure is univocally positive or negative in itself. Consistent with this, we found that outdegree centrality exhibits a curvilinear relationship with desirability, such that positive traits' outdegree centrality is positively associated with desirability,  $r(146) = .46$ , [.32, .57],  $p < .001$ , whereas negative traits' outdegree centrality is negatively associated with desirability,  $r(146) = -.16$ , [-.32, .00],  $p < .05$ . We found that breadth (i.e., a normative measure of informativeness) is not significantly associated with positive,  $r(146) = -.12$ , [-.28, .04],  $p = .147$ , or negative,  $r(146) = -.06$ , [-.22, .10],  $p = .441$ , outdegree centrality, while prevalence is significantly associated with positive,  $r(146) = .29$ , [.14, .43],  $p < .001$ , and marginally with negative,  $r(146) = .16$ , [-.01, .31],  $p = .059$ , outdegree centrality (for full table of trait correlation collapsed across valence's correlations of centrality measures with trait data, see Supplemental Table 4; for full correlation table separated by positive traits, see Supplemental Table 5; for full correlation table separated by negative traits, see Supplemental Table 6).

Finally, we also wanted to distinguish our outdegree measure from another possible network measure of "trait informativeness," to more distinctly isolate the role of number of dependencies in self-evaluations. Here, trait informativeness relates to how much one can infer about other traits based on knowing the value of a given trait. It is formally defined as the reduction in Shannon entropy (Shannon, 1948), an information theoretic measure of uncertainty, measuring the information added (i.e., uncertainty decreased) among all other

traits by including a given trait's ratings (see Supplemental Text for more details). We found that informativeness was correlated with outdegree for positive,  $r(146) = .59$ , [.47, .68], but not negative traits,  $r(146) = .07$ , [-.09, .23], reflecting that while outdegree centrality may be related to informativeness (i.e., uncertainty reduction), they are not identical.

## Discussion

In Study 1, we constructed a trait dependency network and generated predictions from this network to be used for Studies 2 and 3. We tested and validated network reliability and stability across several metrics. First, the thresholding simulations demonstrate the stability of the network across a range of thresholds. Second, a bootstrap procedure that iteratively resampled the network over participants suggests that the network measures have good reliability. While we cannot fully eliminate all concerns about the generalizability of the network under different samples, the stability and reliability tests as well as the use of consensus in achieving a normative network suggest a relatively robust and generalizable network that should reflect normative beliefs about trait dependencies.

In addition, we found that our network measures were related to some normative trait measures used in the past literature and distinct from others. Notably, desirability was positively related to outdegree for positive traits, and negatively related to outdegree centrality for negative traits. However, it is important to note that desirability is distinct from outdegree centrality. Outdegree centrality was computed solely from participants' endorsements of dependencies, and it is doubtful that being seen as socially desirable would make people endorse more dependencies for a given trait on its own. Traits with more dependencies (i.e., higher outdegree centrality) should be more socially desirable if they are positive and less socially desirable if they are negative, because of their number of dependencies, rather than social desirability being why they have more dependencies. For example, one may have a greater desire to be "thoughtful" than "conscientious" because one believes it has more implications on other trait beliefs. People may perceive these traits as desirable because they permit favorable evaluations on dependent traits without contradiction. Doing so may help to preserve positivity and coherence, as possessing higher outdegree positive traits means that one can possess positive traits that depend upon these traits without contradiction or loss of coherence between one's self-view and dependent self-views.

Interestingly, although trait breadth is conceptually related to outdegree centrality in some ways, it was not correlated with outdegree centrality here. This is likely because trait breadth explicitly asks people to judge the extent to which traits relate to a variety of contexts and behaviors, whereas our outdegree measure is constructed from ratings of dependency amongst traits. Thus, our findings suggest that there are not strong empirical relationships between the number of traits that are dependent on a given trait and how many contexts or behaviors may involve a trait. Last, our outdegree measure was also related to, but distinct from, a principled information theoretic measure of informativeness, suggesting that outdegree is not simply measuring how much one can infer about oneself given a specific trait.

## Study 2

The goal of Study 2 was to test our predictions that structured beliefs about trait dependencies shape the self-concept, using the trait dependency network to predict self-evaluations on an independent group of participants. We predicted that more trait dependencies would be associated with more favorable self-evaluations. Specifically, we expected participants to rate higher outdegree positive traits as more self-descriptive and higher outdegree negative traits as less self-descriptive. We also tested whether outdegree centrality would predict more consistent self-evaluations (with trait neighbors), and if individuals lower in self-esteem and higher in depressive symptoms would self-evaluate differently as a function of outdegree centrality.

## Method

### Participants

Two hundred ninety-one native English-speaking participants were recruited with consent and in compliance with UC Riverside institutional review board practices via Amazon Mechanical Turk (62.5% male, 37.5% female), from ages 21–65 ( $M_{\text{age}} = 35.77$ ). Participants were 73.9% Caucasian, 9.6% African-American, 7.6% Hispanic, 4.5% Asian, and 3.8% mixed/other. The mechanical turk participants are limited in terms of racial diversity but vary substantially in terms of age range. Mechanical turk may not be fully generalizable to the U.S. population (Chandler et al., 2019), as mechanical turk workers tend to be younger, more liberal, better educated, less religious, and more likely to be single compared to the U.S. population (Huff & Tingley, 2015; Levay et al., 2016). Participants were compensated \$3 for participation.

We aimed to recruit a large enough sample size to detect cross-level interactions between subject-level individual differences and trial-level network measures (outdegree and indegree). Power analyses of cross-level interactions using simulations (Mathieu et al., 2012) across a variety of parameter configurations (e.g., ICC = .05, intercept variance = .10, slope variance = .01, residual variance = .60, Level 1 effect = .17, Level 2 effect = .05, Level 1 intercept =  $-0.05$ , cross-level interaction = .150) using a Level 2 sample of 291 and a Level 1 sample of 296 consistently return above 80% power. Due to the high number of Level 1 units, the sample size is well-powered to detect cross-level interactions.

### Procedure

Participants were randomly administered all 296 trait words from the positive and negative trait networks and were asked, “How [TARGET TRAIT] are you?” from 1 (*not at all*) to 5 (*very much*). Following the self-evaluations for all 296 traits, participants self-reported on several inventories.

### Self-Report Measures

A variety of self-report measures of individual differences were collected. For brevity, we highlight and focus on two individual differences primarily associated with psychological adjustment and well-being, self-esteem and depressive symptoms. However, the other individual differences measures can be found in the Supplemental Materials.

**Rosenberg Self-Esteem Scale.** A 10-item questionnaire that assesses individual differences in global personal self-esteem was administered (Rosenberg, 1989). The scale demonstrated excellent reliability in the current sample ( $\omega = .97$ ).

**Center for Epidemiologic Studies Depression Scale.** A 20-item questionnaire that measures self-reported symptoms related to depression was administered (Radloff, 1977). The scale demonstrated excellent reliability in the current sample ( $\omega = .97$ ).

## Analysis Plan

Mixed models were implemented in R using *lme4* (Bates et al., 2015) and Satterthwaite’s approximation of degrees of freedom was used for determining  $p$  values in *lmerTest* (Kuznetsova et al., 2017). Pseudo- $R^2$  and semipartial  $R^2$  (standardized generalized variance approach) for linear mixed models were estimated using *r2glmm* (Edwards et al., 2008) and *MuMIn* (Barton, 2009). Likelihood ratio tests (LRT) were performed to determine models best supported by the data. Maximal random effects were tested and were removed as needed if unsupported by the data (i.e., low variance estimates) or if the model failed to converge (Barr et al., 2013).

Trial-level variables of valence, outdegree centrality, indegree centrality, and the interaction of centrality measures with valence were tested as fixed effects, with participants modeled as random factors. Valence was set as a random slope while outdegree centrality and indegree centrality were set as fixed because the variance components for the centrality measures were not supported by the data, and random slopes resulted in a singular fit. Including both indegree and outdegree centrality in the model allowed us to test unique effects of directionality beyond the effect of total number of connections. Due to a typographical error in generating the task, one positive and one negative trait were missing from the task and thus are not included in analyses for Study 2.

We additionally tested how pairwise similarity predicts similarity of responses. Every possible pairwise combination of traits for each valence was extracted,  $C(148, 2) = 10,878$ , and across all participants the same set of pairwise similarity measures were applied to each trait pair while within each participant the distance (absolute value of the difference) between each pair of traits was computed. Therefore, there were  $10,878 * 2 = 21,756$  pairs nested within each participant where the similarities were constant across participants and the distances differed across participants (given that self-evaluations differed between participants, distances in self-evaluations also differed). A mixed model was conducted for these pairs of traits, including the similarity and valence of trait pairs nested within subjects as fixed effects and the distance as the response variable. The interaction of pairwise similarity with valence was tested, as well. All effects in the pairwise analysis were set with fixed slopes.

As a secondary analysis, we tested how centrality predicts network inconsistency. The distance between a trait self-evaluation and the average of its neighbors was computed for all traits, as a measure of network inconsistency (i.e., the greater the distance between a trait self-evaluation and the self-evaluations of its neighbors, the more inconsistent). Outdegree centrality and indegree centrality predicted network inconsistency, with a random slope modeled for outdegree centrality. A random slope for indegree centrality was not supported.



For the models testing centrality's effect on self-evaluations, we tested for interactions between self-esteem and depressive symptoms with indegree centrality and valence as well as with outdegree centrality and valence, modeling three-way interactions between the individual difference measure, valence, and the centrality measure in separate models. To test the effect of degree centrality on network inconsistency, additional models were tested for an interaction of the individual differences with indegree centrality and outdegree centrality.

We tested additional models to determine whether outdegree centrality is predictive of self-evaluations and behavior, while controlling for other normative characteristics such as breadth, prevalence, desirability, interpersonal, and observability. We tested seven variants of the favorability model (outdegree interacting with valence to predict self-evaluations) in which each normative characteristic is controlled for, as well as informativeness and the interaction of informativeness with valence. We employ this same procedure for the consistency model (outdegree predicting inconsistency).

Valence was dummy coded and negative traits were set as the reference level in all analyses. Simple effects of the mixed models were interrogated by analyzing the models with the subset of each valence separately.

### Transparency and Openness

We describe the cleaning and data generating procedures, as well as the analysis approach and modeling decisions. We report the motivation for the sample size and describe all measures used in analyses. We thoroughly describe how data was collected and provide sufficient information and materials to reproduce results. Data, analysis code, and research materials are available at [https://osf.io/3pcvt/?view\\_only=28df28d593354c5a9d65194441877a89](https://osf.io/3pcvt/?view_only=28df28d593354c5a9d65194441877a89). Study 2 was not preregistered.

### Results

We first tested the effect of valence on self-evaluations to replicate the well-documented tendency for people to evaluate themselves favorably. Indeed, we found that valence predicted trial-level self-evaluations,  $\beta = .992$ ,  $SE = .043$ ,  $CI [0.907, 1.076]$ ,  $t(290) = 23.071$ ,  $p < .0001$ ,  $sr^2 = .322$ , such that people evaluated positive traits as more descriptive and negative traits as less descriptive.

### Network Centrality Predicts Favorability of Self-Evaluations

We tested our preregistered prediction that people would evaluate themselves more favorably on higher outdegree centrality traits. Specifically, we predicted that outdegree centrality would be associated with greater self-descriptiveness among positive traits and diminished self-descriptiveness among negative traits. We controlled for indegree centrality to ensure that effects were explained above and beyond overall, bidirectional connections. Consistent with our hypothesis, outdegree centrality interacted with valence to predict self-evaluations,  $\beta = .173$ ,  $SE = .005$ ,  $CI [.163, .182]$ ,  $t(84,970) = 35.002$ ,  $p < .0001$ ,  $sr^2 = .014$ , such that the effect of outdegree centrality on self-descriptiveness was significantly

more positive for positive than for negative traits (see Figure 2A for predicted effects; see Supplemental Figure 3A for raw data). The interaction was decomposed into simple effects, which confirmed that outdegree centrality was negatively associated with self-evaluations for negative traits,  $\beta = -.015$ ,  $SE = .004$ ,  $CI [-.023, -.008]$ ,  $t(42,480) = -4.067$ ,  $p < .0001$ ,  $sr^2 = 0.001$ , and positively associated with self-evaluations for positive traits,  $\beta = .188$ ,  $SE = .004$ ,  $CI [.180, .197]$ ,  $t(42,480) = 43.741$ ,  $p < .0001$ ,  $sr^2 = 0.055$ . This suggests that people self-evaluate more favorably on traits with a higher number of dependencies, as these traits have many perceived downstream consequences and thus have greater impact on self-concept positivity and coherence. It is more likely that the dependencies influenced self-evaluations rather than the reverse, as the dependency judgments that form the network were generated from an independent sample of raters. Indegree, a different type of centrality, exhibited a different pattern of results (see Table 1 for full model results; see Supplemental Table 7 for all model parameters and comparisons; see Supplemental Figure 3B for raw data for indegree interaction; see Supplemental Figure 4 for predicted effects of indegree centrality interaction). Models controlling for normative characteristics and informativeness (both individually and all simultaneously) support the robustness and distinct effect of outdegree centrality on self-evaluations (see Supplemental Table 8; see Supplemental Material for favorability model with all covariates modeled simultaneously). These findings highlight how the structure of trait beliefs relates to self-evaluations. People evaluate traits higher in outdegree centrality more favorably, above and beyond overall network connections. People may maintain positivity and coherence within the self-concept structure by evaluating favorably on traits with many perceived dependencies, which may allow people to evaluate favorably on downstream traits.

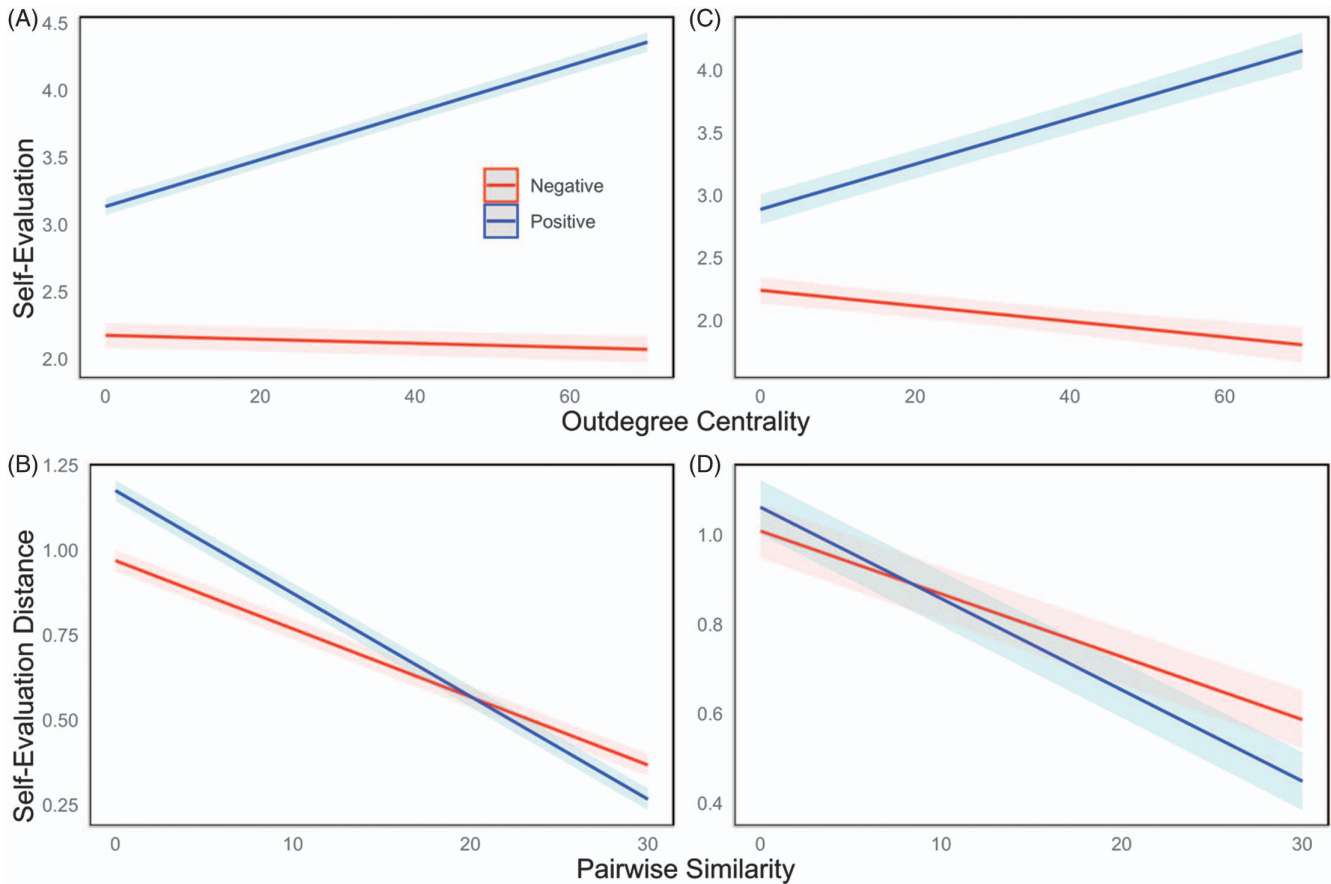
### Network Relationships Predict Consistency of Self-Evaluations

We next examined whether the structural interrelationships between traits relate to self-evaluations. To do so, we tested whether pairwise network similarity (i.e., the number of common neighbors shared by two traits) predicts pairwise differences in self-evaluations. Of primary interest, we first tested a main effect of network similarity on self-evaluation distance. We found a negative effect,  $\beta = -.057$ ,  $SE = .001$ ,  $CI [-.058, -.056]$ ,  $t(6,331,000) = -103.415$ ,  $p < .001$ , suggesting that as the network similarity between a pair of traits increases, the similarity between their self-evaluations increases (Figure 2B). However, we found there was also a significant interaction between network similarity and valence,  $\beta = -.029$ ,  $SE = .001$ ,  $CI [-.031, -.028]$ ,  $t(6,331,000) = -38.17$ ,  $p < .001$ , whereby the effect of network similarity on self-evaluation distance was stronger for positive,  $\beta = -.088$ ,  $SE = .0005$ ,  $CI [-.089, -.087]$ ,  $t(3,165,000) = -162.7$ ,  $p < .001$ , than negative,  $\beta = -.056$ ,  $SE = .0005$ ,  $CI [-.057, -.055]$ ,  $t(3,165,000) = -107.1$ ,  $p < .001$ , traits. While the effect is stronger for positive than negative similarity, network similarity predicts more similarity in evaluations regardless of valence. These findings indicate that traits that are structurally similar (i.e., share many neighbors) are evaluated similarly (see Table 2 for full model results; see Supplemental Table 9 for model statistics).

While the pairwise similarity analysis suggests that traits with overlapping neighbors are evaluated similarly, we also sought to test

**Figure 2**

*Predicted Effects of Outdegree Centrality on Self-Evaluations (Top) and Pairwise Similarity (Bottom) on Self-Evaluation Distance*



*Note.* A and B are study 2 and C and D are study 3. Predicted effects while holding covariates constant, with confidence intervals of  $\pm 1.96$  SE. (A/C) The interaction of outdegree centrality and valence on self-evaluations. Outdegree centrality predicts favorability in self-evaluations. That is, positive traits were evaluated as more self-descriptive with increases in outdegree centrality and negative traits were evaluated as less self-descriptive with increases in outdegree centrality. (B/D) The effect of pairwise similarity on self-evaluation distance, such that traits with greater network-defined similarity are evaluated more similarly (stronger for positive than for negative pairs).

whether outdegree centrality influences how consistent a trait self-evaluation is with its neighboring traits' self-evaluations. Specifically, we computed the distance between the self-evaluation for a trait and the average self-evaluation of its immediate, first-order neighbors as a measure of network inconsistency (i.e., larger distance reflects greater inconsistency with neighbors). We found that outdegree centrality, while controlling for indegree centrality,  $\beta = -.070$ ,  $SE = .004$ ,  $CI [-.077, -.062]$ ,  $t(85,260) = -17.071$ ,  $p < .0001$ ,  $sr^2 = .005$ , was strongly negatively associated with inconsistency, such that as outdegree centrality increases, consistency in self-evaluations between a trait and its neighbors increases (see Table 3 for full model results; see Supplemental Table 10 for model parameters and comparisons). Models controlling for normative characteristics and informativeness support robustness and unique effects of outdegree centrality on self-evaluation consistency (see Supplemental Table 11; see Supplemental Material for consistency with all covariates modeled simultaneously). These findings provide additional support for outdegree centrality as important to maintaining a locally consistent, and thereby globally coherent, self-concept.

### **Individual Differences Moderate Centrality Effects on Self-Evaluations**

As a further test of our network model of self-representation, we examined how individual differences in self-esteem and depressive symptomatology moderated the effect of outdegree on the favorability and consistency of self-evaluations. We first examined if individual differences in self-esteem and depressive symptoms moderated the valence-dependent outdegree centrality effects on favorable evaluations. We found a three-way interaction for valence, outdegree, and self-esteem,  $\beta = .025$ ,  $SE = .005$ ,  $CI [.016, .035]$ ,  $t(85,550) = 5.139$ ,  $p < .0001$ ,  $sr^2 = .0003$ , suggesting that individuals higher in self-esteem differentiated outdegree centrality in their self-evaluations more than those lower in self-esteem, for positive but not negative traits. We also identified a consistent three-way interaction of valence and outdegree centrality with depressive symptoms,  $\beta = -.047$ ,  $SE = .005$ ,  $CI [-.056, -.037]$ ,  $t(85,550) = -9.507$ ,  $p < .0001$ ,  $sr^2 = .001$ . Inspection of the simple effects revealed that individuals higher in self-esteem,  $\beta = .025$ ,  $SE = .004$ ,  $CI [.017, .034]$ ,  $t(42,780) = 5.915$ ,

**Table 1***Mixed Model Table of Outdegree Interacting With Valence Predicting Self-Evaluations for Study 2 and Study 3*

Fixed effects	Study 2			Study 3		
	$\beta$ (SE)	$t$	$p$	$\beta$ (SE)	$t$	$p$
Intercept	−0.496 (.034)	−14.750	<.001	−0.506 (.041)	−12.280	<.001
Outdegree	−0.014 (.003)	−3.918	<.001	−0.062 (.011)	−5.556	<.001
Valence (Pos)	0.992 (.043)	23.071	<.001	1.011 (.076)	13.260	<.001
Indegree	−0.036 (.003)	−11.439	<.001	−0.046 (.009)	−5.025	<.001
Outdegree $\times$ Valence (Pos)	0.173 (.005)	35.002	<.001	0.241 (.014)	16.677	<.001
Indegree $\times$ Valence (Pos)	0.031 (.005)	6.133	<.001	0.009 (.015)	0.589	.556
Random effects						
$\sigma^2$	0.50			0.63		
$\tau_{00}$	0.33 <sub>Subject</sub>			0.07 <sub>Subject</sub>		
$\tau_{11}$	0.53 <sub>Subject.Positive</sub>			0.25 <sub>Subject.Positive</sub>		
$\rho_{01}$	−0.85 <sub>Subject</sub>			−0.79 <sub>Subject</sub>		
ICC	0.32			0.13		
$N$	291 <sub>Subject</sub>			45 <sub>Subject</sub>		
Observations	85,554			12,876		
Marginal $R^2$ /conditional $R^2$	0.260/0.497			0.277/0.368		

*Note.* Negative is the reference group for the dummy-coded valence variable. Parameter slopes are all fixed except for valence.  $\sigma^2$  represents the within-subject variance (residual),  $\tau_{00}$  represents the between-subject variance (random intercept for subject),  $\tau_{11}$  represents the positive valence variance (random slope for valence),  $\rho_{01}$  represents the random correlation between subject's intercept and subject's valence slope. ICC is the intraclass correlation coefficient.  $N$  depicts the number of Level-2 subject units and observations depicts the total number of units across all trials. Marginal  $R^2$  is the pseudo  $R$ -squared estimate for fixed effects and marginal  $R^2$  is the pseudo  $R$ -squared estimate for fixed and random effects. Pos = positive.

$p < .0001$ ,  $sr^2 = .001$ , and lower in depressive symptoms,  $\beta = -.048$ ,  $SE = .004$ ,  $CI [-.056, -.039]$ ,  $t(42,780) = -11.117$ ,  $p < .0001$ ,  $sr^2 = .004$ , evaluated positive traits higher in outdegree centrality as more self-descriptive and positive traits lower in outdegree centrality as less self-descriptive, whereas those lower in self-esteem or higher in depressive symptoms did not differentiate the outdegree centrality of traits in their self-evaluations. We did not find that self-esteem,  $\beta = -.004$ ,  $SE = .004$ ,  $CI [-.012, .003]$ ,  $t(42,780) = -1.151$ ,  $p = .250$ , or depressive symptoms,  $\beta = .007$ ,  $SE = .004$ ,  $CI [-.000, .015]$ ,  $t(42,780) = 1.944$ ,  $p = .502$ , interacted with outdegree centrality for negative traits. The results suggest that individuals higher in self-

esteem have positive self-evaluations that are more associated with outdegree centrality than those lower in self-esteem, potentially facilitating a positive self-concept. By evaluating higher outdegree positive traits as more self-descriptive, other traits that depend upon these traits can also be evaluated positively without a loss of coherence. On the other hand, individuals higher in depressive symptoms may not differentiate their self-evaluations as a function of trait dependencies, and may thus not evaluate more favorably on traits which are perceived as having more dependencies. By evaluating higher outdegree positive traits as less self-descriptive, other traits that depend upon these traits cannot be evaluated positively

**Table 2***Mixed Model Table for Pairwise Similarity Predicting Self-Evaluation Distance for Study 2 and Study 3*

Fixed effects	Study 2			Study 3		
	$\beta$ (SE)	$t$	$p$	$\beta$ (SE)	$t$	$p$
Intercept	−0.084 (.016)	−5.411	<.001	−0.018 (.034)	−0.525	.600
Similarity	−0.057 (.001)	−103.415	<.001	−0.044 (.001)	−29.883	<.001
Valence (Pos)	0.167 (.001)	219.233	<.001	0.032 (.002)	15.955	<.001
Similarity $\times$ Valence (Pos)	−0.029 (.001)	−38.166	<.001	−0.020 (.002)	−9.877	<.001
Random effects						
$\sigma^2$	0.92			0.95		
$\tau_{00}$	0.07 <sub>Subject</sub>			0.05 <sub>Subject</sub>		
ICC	0.07			0.05		
$N$	291 <sub>Subject</sub>			45 <sub>Subject</sub>		
Observations	6,330,996			916,401		
Marginal $R^2$ /conditional $R^2$	0.013/0.083			0.003/0.055		

*Note.* Negative is the reference group for the dummy-coded valence variable. Parameter slopes are all fixed.  $\sigma^2$  represents the within-subject variance (residual) and  $\tau_{00}$  represents the between-subject variance (random intercept for subject). ICC is the intraclass correlation coefficient.  $N$  depicts the number of Level-2 subject units and observations depicts the total number of units across all pairs. Marginal  $R^2$  is the pseudo  $R$ -squared estimate for fixed effects and marginal  $R^2$  is the pseudo  $R$ -squared estimate for fixed and random effects. Pos = positive.

**Table 3***Mixed Model Table for Model of Centrality Predicting Self-Evaluation Consistency With Neighbors for Study 2 and Study 3*

Fixed effects	Study 2			Study 3		
	$\beta$ (SE)	$t$	$p$	$\beta$ (SE)	$t$	$p$
Intercept	0.000 (.017)	0.013	.990	−0.000 (.041)	−0.006	.995
Outdegree	−0.070 (.004)	−17.071	<.001	−0.036 (.010)	−3.765	<.001
Indegree	−0.016 (.003)	−5.000	<.001	−0.013 (.009)	−1.477	.140
Random effects						
$\sigma^2$	0.91			0.93		
$\tau_{00}$	0.09 <sub>Subject</sub>			0.07 <sub>Subject</sub>		
$\tau_{11}$	0.00 <sub>Subject.Outdegree</sub>			0.00 <sub>Subject.Outdegree</sub>		
$\rho_{01}$	0.17 <sub>Subject</sub>			−0.13 <sub>Subject</sub>		
ICC	0.09			0.07		
$N$	291 <sub>Subject</sub>			45 <sub>Subject</sub>		
Observations	85,554			12,876		
Marginal $R^2$ /conditional $R^2$	0.006/0.093			0.002/0.075		

*Note.* Parameter slopes are fixed for indegree and random for outdegree.  $\sigma^2$  represents the within-subject variance (residual),  $\tau_{00}$  represents the between-subject variance (random intercept for subject),  $\tau_{11}$  represents the positive valence variance (random slope for outdegree centrality),  $\rho_{01}$  represents the random correlation between subject's intercept and subject's outdegree slope. ICC is the intraclass correlation coefficient.  $N$  depicts the number of Level-2 subject units and observations depicts the total number of units across all trials. Marginal  $R^2$  is the pseudo  $R$ -squared estimate for fixed effects and marginal  $R^2$  is the pseudo  $R$ -squared estimate for fixed and random effects.

without a loss of coherence, which could reduce the positivity and coherence of the self-concept (see Table 4 for full self-esteem model results; see Table 5 for full depressive symptoms model results; see Supplemental Material for self-esteem and depressive symptoms results with covariates; see Supplemental Figure 5A for self-esteem's predicted effects on favorable self-evaluations and Supplemental Figure 6A for depressive symptoms' predicted effects on favorable self-evaluations).

We additionally tested whether self-esteem and depressive symptoms moderate the effect of outdegree centrality on consistency in self-evaluations. We identified a negative effect of self-esteem,  $\beta = -.087$ ,  $SE = .017$ ,  $CI [-.119, -.054]$ ,  $t(291) = -5.186$ ,  $p < .0001$ ,  $sr^2 = .008$ , and a positive effect of depressive symptoms,  $\beta = .110$ ,  $SE = .016$ ,  $CI [.078, .142]$ ,  $t(291) = 6.804$ ,  $p < .0001$ ,  $sr^2 = .013$ , on network inconsistency, such that individuals higher in self-esteem and lower in depressive symptoms self-evaluated more consistently with trait neighbors. Moreover, self-esteem,  $\beta = -.013$ ,  $SE = .003$ ,  $CI [-.021, -.005]$ ,  $t(85,260) = -3.282$ ,  $p < .0001$ ,  $sr^2 = .0002$ , and depressive symptoms,  $\beta = .014$ ,  $SE = .003$ ,  $CI [.006, .022]$ ,  $t(291) = 3.537$ ,  $p < .0001$ ,  $sr^2 = .0002$ , moderated the effect of outdegree centrality on self-evaluative consistency between traits and their neighbors. Individuals higher in self-esteem and lower in depressive symptoms had more consistent self-evaluations between traits and their neighbors for higher outdegree centrality traits than lower outdegree centrality traits, whereas those lower in self-esteem and higher in depressive symptoms evaluated less consistently with trait neighbors, regardless of outdegree centrality. Thus, individuals higher in psychological adjustment may evaluate most consistently on traits perceived to have many dependencies in order to avoid self-contradictions and a loss of coherence that could result from inconsistent evaluations. Taken together with the above findings, these results are consistent with our hypothesis that more favorable and consistent evaluations on higher outdegree traits help to maintain self-concept positivity and coherence (see Table 6 for full self-esteem model results; see Table 7 for full depressive symptoms

model results; see Supplemental Material for self-esteem and depressive symptoms results with covariates; see Supplemental Figure 5B for self-esteem's predicted effects on consistency and Supplemental Figure 6B for depressive symptoms' predicted effects on consistency).

In addition to depressive symptoms and self-esteem, we tested other individual differences related to personality, self-construal, and well-being on outdegree and indegree centrality effects. These tests demonstrate similar effects across measures but for the purposes of brevity, they can be found in Supplemental Materials (see Supplemental Table 12 for model statistics of all individual differences for outdegree effect on favorable self-evaluations; see Supplemental Table 13 for model statistics of all individual differences for outdegree effect on consistency).

## Discussion

In Study 2, we provide initial support for the trait dependency network developed in Study 1. Supporting our primary hypothesis, outdegree centrality is associated with more favorable and consistent self-evaluations. These findings reflect outdegree centrality's importance to positivity and coherence, as people self-evaluate favorably and maintain more consistency amongst the neighbors of traits that are perceived as having more consequences for other traits. Together, these results are consistent with our overall hypothesis that people maintain a positive and coherent self-concept by favoring traits that are perceived as having more dependencies. By evaluating higher outdegree traits more favorably (i.e., self-descriptive for positive traits and nonself-descriptive for negative traits), people avoid contradictions with favorable evaluations on their dependent traits, which simultaneously preserves coherence (by not having contradictory self-views) and overall positivity. Additionally, we found that self-esteem and depressive symptoms moderated our behavioral effects such that outdegree centrality exhibited stronger effects for individuals



**Table 4***Mixed Model Table for Self-Esteem Moderating Outdegree and Valence Effects on Self-Evaluations for Study 2 and Study 3*

Fixed effects	Study 2			Study 3		
	$\beta$ (SE)	<i>t</i>	<i>p</i>	$\beta$ (SE)	<i>t</i>	<i>p</i>
Intercept	−0.496 (.030)	−16.283	<.001	−0.506 (.037)	−13.569	<.001
Outdegree	−0.014 (.003)	−3.919	<.001	−0.062 (.011)	−5.578	<.001
Valence (Pos)	0.992 (.036)	27.247	<.001	1.011 (.055)	18.465	<.001
Self-esteem (SE)	−0.244 (.030)	−8.010	<.001	−0.122 (.037)	−3.273	.001
Indegree	−0.036 (.003)	−11.442	<.001	−0.046 (.009)	−4.998	<.001
Outdegree × Positive	0.173 (.005)	35.012	<.001	0.242 (.014)	16.700	<.001
SE × Outdegree	−0.004 (.003)	−1.109	.268	0.016 (.011)	1.455	.146
SE × Positive	0.391 (.036)	10.734	<.001	0.356 (.055)	6.505	<.001
SE × Indegree	−0.010 (.003)	−3.095	.002	−0.017 (.009)	−1.919	.055
Indegree × Positive	0.031 (.005)	6.135	<.001	0.008 (.015)	0.574	.566
Outdegree × SE × Positive	0.025 (.005)	5.139	<.001	0.007 (.014)	0.483	.629
Indegree × SE × Positive	0.012 (.005)	2.414	.016	0.002 (.015)	0.147	.883
Random effects						
$\sigma^2$	0.50			0.63		
$\tau_{00}$	0.27 <sub>Subject</sub>			0.06 <sub>Subject</sub>		
$\tau_{11}$	0.38 <sub>Subject,Positive</sub>			0.13 <sub>Subject,Positive</sub>		
$\rho_{01}$	−0.81 <sub>Subject</sub>			−0.75 <sub>Subject</sub>		
ICC	0.28			0.08		
<i>N</i>	291 <sub>Subject</sub>			45 <sub>Subject</sub>		
Observations	85,554			12,876		
Marginal $R^2$ /conditional $R^2$	0.301/0.498			0.313/0.370		

*Note.* Negative is the reference group for the dummy-coded valence variable. Parameter slopes are all fixed except for valence.  $\sigma^2$  represents the within-subject variance (residual),  $\tau_{00}$  represents the between-subject variance (random intercept for subject),  $\tau_{11}$  represents the positive valence variance (random slope for valence),  $\rho_{01}$  represents the random correlation between subject's intercept and subject's valence slope. ICC is the intraclass correlation coefficient. *N* depicts the number of Level-2 subject units and observations depicts the total number of units across all trials. Marginal  $R^2$  is the pseudo *R*-squared estimate for fixed effects and marginal  $R^2$  is the pseudo *R*-squared estimate for fixed and random effects. Pos = positive; SE = self-esteem.

higher in self-esteem and lower in depressive symptoms. We suggest that individuals higher in self-esteem may be more attuned to trait dependencies, which allows them to better maintain self-concept positivity and coherence. Conversely, individuals higher in depressive symptoms may be less effective at discriminating normative beliefs about trait dependencies, which may impede the maintenance of positivity and coherence within the self-concept. These results are correlational, so it is impossible to say whether depressive symptoms lead to less attention to traits with more dependencies or whether attending less to traits with more dependencies leads to depressive symptoms. Further, the results could also reflect the possibility that depressive symptoms are associated with different normative beliefs about what traits are more central than others. We believe it is most likely that not discriminating between perceived dependency relations may limit an individual's ability to transmit positivity and maintain noncontradictory self-beliefs across trait evaluations, which may contribute to mental health dysfunction and maladjustment.

Lastly, while in Study 1 we correlated outdegree centrality with normative characteristics and network-derived informativeness to test how discriminable they are, in Study 2, we performed additional models where we controlled for these variables in our favorability, consistency, and individual differences moderated models. Outdegree effects remain largely unchanged, with the exception of network-derived informativeness as a covariate in the consistency model. This is unsurprising as informativeness incorporates similarity and self-evaluation information into its estimation, and thus

carries more information about its neighbors' and their evaluations than outdegree alone.

### Study 3

The goal of Study 3 was to test whether normative beliefs about trait dependencies are reflected in brain processing during self-reflection. Participants completed the same task as in Study 2, this time while undergoing a functional magnetic resonance imaging (fMRI) task. First, we test the same behavioral findings revealed in Study 2 in this new sample. Second, we test whether whole-brain activation during self-evaluation is associated with outdegree centrality, with an *a priori* focus on vmPFC given its involvement in related process, such as self, trait, and concept processing. Finally, we test whether traits that have similar dependency relations in the trait dependency network elicit similar activation patterns.

### Method

#### Participants

An independent sample of 45 undergraduate student participants (31.1% male, 68.9% female) from ages 17–26 ( $M_{\text{age}} = 19.47$ ) were recruited with consent and in compliance with approved institutional review board practices and via the UC Riverside SONA credit system. They were compensated two credits for their participation. Participants identified as 37.8% Asian, 31.1% Hispanic, 8.9% Caucasian, 8.9% mixed, 6.7% other, 4.4% African-American, and 2.2%

**Table 5***Mixed Model Table for Depressive Symptoms Moderating Outdegree and Valence Effects on Self-Evaluations for Study 2 and Study 3*

Fixed effects	Study 2			Study 3		
	$\beta$ (SE)	<i>t</i>	<i>p</i>	$\beta$ (SE)	<i>t</i>	<i>p</i>
Intercept	−0.496 (.022)	−22.530	<.001	−0.504 (.037)	−13.518	<.001
Outdegree	−0.014 (.003)	−3.922	<.001	−0.061 (.011)	−5.433	<.001
Valence (Pos)	0.992 (.030)	33.593	<.001	1.022 (.066)	15.400	<.001
Depression (Dep)	0.432 (.022)	19.654	<.001	0.134 (.037)	3.585	<.001
Indegree	−0.036 (.003)	−11.451	<.001	−0.047 (.009)	−5.094	<.001
Outdegree × Positive	0.173 (.005)	35.039	<.001	0.237 (.015)	16.263	<.001
Dep × Outdegree	0.006 (.003)	1.874	.061	−0.016 (.011)	−1.428	.153
Dep × Positive	−0.532 (.030)	−18.029	<.001	−0.267 (.066)	−4.028	<.001
Dep × Indegree	0.031 (.005)	6.140	<.001	0.011 (.015)	0.763	.446
Indegree × Positive	0.016 (.003)	5.071	<.001	0.011 (.009)	1.211	.226
Outdegree × Dep × Positive	−0.047 (.005)	−9.507	<.001	−0.011 (.015)	−0.759	.448
Indegree × Dep × Positive	−0.025 (.005)	−4.892	<.001	−0.011 (.015)	−0.736	.462
Random effects						
$\sigma^2$	0.50			0.62		
$\tau_{00}$	0.14 <sub>Subject</sub>			0.06 <sub>Subject</sub>		
$\tau_{11}$	0.25 <sub>Subject.Positive</sub>			0.18 <sub>Subject.Positive</sub>		
$\rho_{01}$	−0.67 <sub>Subject</sub>			−0.74 <sub>Subject</sub>		
ICC	0.22			0.11		
<i>N</i>	291 <sub>Subject</sub>			44 <sub>Subject</sub>		
Observations	85,554			12,584		
Marginal $R^2$ /conditional $R^2$	0.360/0.498			0.302/0.376		

*Note.* Different inventories measured depressive symptoms in Study 2 (Center for Epidemiological Studies-Depression) and Study 3 (Beck Depression Inventory). Negative is the reference group for the dummy-coded valence variable. Parameter slopes are all fixed except for valence.  $\sigma^2$  represents the within-subject variance (residual),  $\tau_{00}$  represents the between-subject variance (random intercept for subject),  $\tau_{11}$  represents the positive valence variance (random slope for valence),  $\rho_{01}$  represents the random correlation between subject's intercept and subject's valence slope. ICC is the intraclass correlation coefficient. *N* depicts the number of Level-2 subject units and observations depicts the total number of units across all trials. Marginal  $R^2$  is the pseudo *R*-squared estimate for fixed effects and marginal  $R^2$  is the pseudo *R*-squared estimate for fixed and random effects. Pos = positive; Dep = depression.

Pacific Islander. The participant sample is a racially diverse undergraduate sample that extends beyond the Amazon Mechanical Turk sample used in Study 2. While the sample used in Study 2 was predominantly White but sampled from a wide range of ages, the sample in Study 3 was more limited in age but generalizes across broader racial demographics. Notably, while Study 2 undersampled from Hispanic populations, Study 3 sampled heavily from Hispanic and Asian populations. Across both studies, participants were sampled across a relatively broad range of racial and age demographics.

Participants were screened for any potential contraindications or any protocol-related exclusion criteria, and were all right-handed, native English speakers, free from medications, and psychological and neurological conditions, and had normal or corrected-to-normal vision. The present study's sample size doubles that of a previously reported pilot study (Beer & Hughes, 2010 for original data; Davis et al., 2014 for network reanalysis), which the present study replicates and extends. While we did not conduct a formal power analysis for the fMRI analysis, a simulated power analysis (Green & MacLeod, 2016) on the behavioral findings from the primary within-subjects interaction between outdegree centrality and valence in Study 2 revealed that power above 90% is achieved at Study 3's sample size.

### Behavioral Procedure

**Task.** Participants underwent a standard self-evaluative fMRI task (e.g., Moran et al., 2006). Participants completed four functional runs in which they were asked to evaluate themselves on 296 trait

words across runs—148 positive traits and 148 negative traits from the trait dependency networks. On each trial, participants were asked, “To what extent does the following trait describe you?” Below the question and trait word, an ordinal Likert scale, ranging from 1 (*not at all*) to 5 (*extremely*) was displayed for participants to self-evaluate on. Participants were able to respond for up to 3 s per trial. Once a response was selected, responses were highlighted in orange until the trial had ended. Each trial was followed by a jittered intertrial interval. For each run, we drew random numbers from a truncated exponential distribution with a minimum of 2 s and a mean of 3 s. The task was presented using MATLAB's Psychtoolbox and projected onto a screen that was viewed via a mirror mounted on the scanner.

### Self-Report Measures

**Rosenberg Self-Esteem Scale.** A 10-item questionnaire that assesses individual differences in global personal self-esteem was administered (Rosenberg, 1989). The scale demonstrated excellent reliability in the current sample ( $\omega = .94$ ).

**Beck Depression Inventory.** A 21-item questionnaire that measures self-reported symptoms related to depression was administered (Beck et al., 1987). The scale demonstrated excellent reliability in the current sample ( $\omega = .92$ ).

### Analysis Plan

**Behavioral Analysis.** Identical mixed models were run as in Study 2.

**Table 6***Mixed Model Table for Self-Esteem Moderating Effects of Centrality on Self-Evaluation Consistency With Neighbors for Study 2 and Study 3*

Fixed effects	Study 2			Study 3		
	$\beta$ (SE)	$t$	$p$	$\beta$ (SE)	$t$	$p$
Intercept	0.000 (.017)	0.013	.989	−0.000 (.041)	−0.006	.995
Outdegree	−0.070 (.004)	−17.440	<.001	−0.036 (.009)	−3.915	<.001
Self-esteem (SE)	−0.087 (.017)	−5.186	<.001	−0.042 (.041)	−1.016	.310
Indegree	−0.016 (.003)	−5.000	<.001	−0.013 (.009)	−1.480	.139
Outdegree $\times$ SE	−0.013 (.004)	−3.282	.001	−0.020 (.009)	−2.199	.028
Indegree $\times$ SE	−0.010 (.003)	−3.000	.003	0.001 (.009)	0.135	.893
Random effects						
$\sigma^2$	0.91			0.93		
$\tau_{00}$	0.08 <sub>Subject</sub>			0.07 <sub>Subject</sub>		
$\tau_{11}$	0.00 <sub>Subject, Outdegree</sub>			0.00 <sub>Subject, Outdegree</sub>		
$\rho_{01}$	0.08 <sub>Subject</sub>			−0.32 <sub>Subject</sub>		
ICC	0.08			0.07		
$N$	291 <sub>Subject</sub>			45 <sub>Subject</sub>		
Observations	85,554			12,876		
Marginal $R^2$ /conditional $R^2$	0.013/0.093			0.004/0.077		

*Note.* Parameter slopes are all fixed except for outdegree centrality.  $\sigma^2$  represents the within-subject variance (residual),  $\tau_{00}$  represents the between-subject variance (random intercept for subject),  $\tau_{11}$  represents the positive valence variance (random slope for outdegree centrality),  $\rho_{01}$  represents the random correlation between subject's intercept and subject's outdegree slope. ICC is the intraclass correlation coefficient.  $N$  depicts the number of Level-2 subject units and observations depicts the total number of units across all trials. Marginal  $R^2$  is the pseudo  $R$ -squared estimate for fixed effects and marginal  $R^2$  is the pseudo  $R$ -squared estimate for fixed and random effects. SE = self-esteem.

**Imaging Acquisition.** Imaging data were acquired on a 3T Magnetic Resonance Imaging (MRI) scanner (Prisma, Siemens Healthineers, Malvern, PA) at the UC Riverside Center for Advanced Neuroimaging using a 32-channel receive only coil. Images from a T<sub>1</sub>-weighted Magnetization Prepared - Rapid Gradient Echo (MP-RAGE) sequence (echo time [TE]/repetition time [TR]/inversion time = 3.02 ms/2,600 ms/800 ms, flip angle [FA] = 8°, voxel size = 0.8

$\times 0.8 \times 0.8$  mm<sup>3</sup>) were used to position imaging volumes in functional scans in addition to use for registration from subject space to common space.

Functional data were collected with an T<sub>2</sub>\*-weighted gradient echo planar imaging (EPI) sequence with the following scan parameters: TE/TR = 32 ms/1,700 ms; slices = 72; FA = 75°, field-of-view; FOV = 220 mm  $\times$  190 mm; matrix size = 130 $\times$ 112;

**Table 7***Mixed Model Table for Depressive Symptoms Moderating Effects of Centrality on Self-Evaluation Consistency With Neighbors for Study 2 and Study 3*

Fixed effects	Study 2			Study 3		
	$\beta$ (SE)	$t$	$p$	$\beta$ (SE)	$t$	$p$
Intercept	0.000 (.016)	0.014	.989	−0.007 (.042)	−0.157	.875
Outdegree	−0.070 (.004)	−17.506	<.001	−0.036 (.010)	−3.758	<.001
Depression (Dep)	0.110 (.016)	6.781	<.001	0.039 (.042)	0.940	.347
Indegree	−0.016 (.003)	−5.000	<.001	−0.013 (.009)	−1.492	.136
Outdegree $\times$ Dep	0.014 (.004)	3.537	<.001	0.019 (.010)	1.912	.056
Indegree $\times$ Dep	0.011 (.003)	3.312	.001	−0.004 (.009)	−0.451	.652
Random effects						
$\sigma^2$	0.91			0.92		
$\tau_{00}$	0.07 <sub>Subject</sub>			0.07 <sub>Subject</sub>		
$\tau_{11}$	0.00 <sub>Subject, Outdegree</sub>			0.00 <sub>Subject, Outdegree</sub>		
$\rho_{01}$	0.03 <sub>Subject</sub>			−0.25 <sub>Subject</sub>		
ICC	0.08			0.07		
$N$	291 <sub>Subject</sub>			44 <sub>Subject</sub>		
Observations	85,554			12,584		
Marginal $R^2$ /conditional $R^2$	0.018/0.093			0.003/0.077		

*Note.* Different inventories measured depressive symptoms in Study 2 (Center for Epidemiological Studies-Depression) and Study 3 (Beck Depression Inventory). Parameter slopes are all fixed except for outdegree centrality.  $\sigma^2$  represents the within-subject variance (residual),  $\tau_{00}$  represents the between-subject variance (random intercept for subject),  $\tau_{11}$  represents the positive valence variance (random slope for outdegree centrality),  $\rho_{01}$  represents the random correlation between subject's intercept and subject's outdegree slope. ICC is the intraclass correlation coefficient.  $N$  depicts the number of Level-2 subject units and observations depicts the total number of units across all trials. Marginal  $R^2$  is the pseudo  $R$ -squared estimate for fixed effects and marginal  $R^2$  is the pseudo  $R$ -squared estimate for fixed and random effects. Dep = depression.

voxel size =  $1.7 \times 1.7 \times 1.7 \text{ mm}^3$ ; generalized autocalibrating partially parallel acquisitions (GRAPPA) = 2; multiband factor = 3; bandwidth = 1,540 Hz/pixel, phase encode = AP. A pair of spin echo EPI acquisitions with identical spatial parameters and bandwidth but opposite phase encoding directions (AP and PA) were collected to correct for susceptibility-related distortions.

**Imaging Analysis.** fMRI data preprocessing and statistical analyses were conducted using FEAT (FMRI Expert Analysis Tool) Version 6.00, part of the Functional Magnetic Resonance Imaging of the Brain Software Library (FSL) (Jenkinson et al., 2012). Brain extraction was performed using the Brain Extraction Tool (BET). From the reversed phase-encoded pairs, the susceptibility-induced off-resonance field was estimated using FSL's top up tool (Andersson et al., 2003) and the two images were combined into a single corrected one. Registration of the functional data to the high-resolution structural data was carried out using the boundary-based registration algorithm (Greve & Fischl, 2009). Registration of the high-resolution structural image to standard space (Montreal Neurological Institute template [MNI]) was completed using the FMRIB's Linear Image Registration Tool (FLIRT) (Jenkinson et al., 2002), and was further refined using a 12 degree-of-freedom FMRIB's Non-linear Image Registration Tool (FNIRT) at a warp resolution of 10 mm (Andersson et al., 2007). Partial smoothing was applied using a 6-mm Filtered White Gaussian Noise (FWGN) kernel. The entire 4D data set was grand-mean intensity normalized by a single multiplicative factor. High-pass temporal filtering was applied to remove low frequencies (100 s cut off; Gaussian-weighted least-squares straight line fitting, with  $\sigma = 50.0 \text{ s}$ ).

Time-series statistical analysis was carried out using FMRIB's Improved Linear Model (FILM) with local autocorrelation correction (Woolrich et al., 2001). Statistical analyses were carried out using a standard three-level analysis in FEAT. The first-level model included separate explanatory variables (EVs; continuous variables centered and scaled within subjects) for self-evaluations, outdegree centrality, indegree centrality, and missing responses. A constant EV at each onset was also included to model overall effect of stimulus presentation. Nuisance regressors were included for motion (six motion parameters and their temporal derivatives) and volumes exceeding head motion of .9 mm (Siegel et al., 2014). The model also included temporal derivatives for each task variable. Regressors were convolved to a double-gamma Hemodynamic Response Function (HRF). The second-level model, averaging contrast estimates within subjects, was tested using a fixed effects analysis that forces random effect variance to zero. Finally, a third-level model, averaging contrast estimates between subjects,<sup>1</sup> was tested using FMRIB's local analysis of mixed effects (FLAME) Stage 1, which accounts for both within- and between-subjects variances (Beckmann et al., 2003; Woolrich, 2008; Woolrich et al., 2004). Final statistical maps were corrected for multiple comparisons at  $p < .05$  using permutation-based cluster mass thresholding, implemented in FSL's randomize. This analysis included a primary cluster-forming threshold of  $t = 3.28$  (critical value of  $t$  for  $df = 45$  and  $\alpha = 0.001$ ) and 6 mm variance smoothing.

In addition to the above model, we also tested a supplemental model that incorporated additional EVs as covariates for normative desirability and network-derived informativeness, to account for potential alternative explanations of outdegree centrality's effect on activation.

**Representational Similarity Analysis.** To determine trial-by-trial estimates of the hemodynamic response, we used a Least

Squares - All; LS-A procedure (Mumford et al., 2012) to compute a  $\beta$ -map (Rissman et al., 2004) for each stimulus onset. This involves modeling individual task trials as separate regressors in a single general linear model. The estimated activation patterns for each onset were then registered to standard space through FNIRT. Representational similarity analysis (RSA) was employed to examine how semantic similarity of traits, as determined by the network of dependent relationships between traits, may relate to patterns of neural similarity across the brain.

An RSA searchlight was conducted using MultiVariate Pattern Analysis in Python; PyMVPA2 (Hanke et al., 2009). A similarity matrix between traits was determined by the inverse logarithm weighted similarity measures produced from network. Similarity of trial-level parameter estimates were computed for each pair of traits with Pearson correlations. The spatial localization of voxels within the statistical map was determined by applying a searchlight algorithm (Kriegeskorte et al., 2006) with a 5-voxel radius. The searchlight iterated across each subject's brain and computed correlations between the lower triangle of the neural response similarity matrix and the lower triangle of the pairwise trait similarity matrix for each valence and stored the correlation coefficient at each voxel at the center of searchlight. Upon completion of the searchlight for each valence, the positive and negative valence RSA statistical maps were concatenated across subjects and averaged between positive and negative valence. Subsequently, using the concatenated and averaged RSA map, correlations were transformed to  $t$  values and were tested at the group-level using permutation-based thresholding, at a primary cluster-forming threshold consistent with the univariate analyses.

### Transparency and Openness

We describe the preprocessing, cleaning, and data generating procedures, as well as the analysis approach, and modeling decisions. We report the motivation for the sample size, and describe all measures used in analyses. We thoroughly describe how data was collected and provide sufficient information and materials to reproduce results. Data, analysis code, and research materials for both the behavioral and the neuroimaging analyses are available at [https://osf.io/3pcvt/?view\\_only=28df28d593354c5a9d65194441877a89](https://osf.io/3pcvt/?view_only=28df28d593354c5a9d65194441877a89). The neuroimaging statistical maps are located at <https://neurovault.org/collections/CRTMNQFW/>. The hypotheses were preregistered for Study 3 and not Study 2, and the preregistration is located at <http://aspredicted.org/blind.php?x=89ag6d>. Study 2 appears before Study 3 narratively in the current text, but chronologically, we collected data for Study 2 after Study 3. We preregistered the primary findings regarding vmPFC response tracking outdegree centrality and more favorable self-evaluations being associated with higher outdegree centrality. The individual differences analyses were largely exploratory in Study 3, but were approached in a confirmatory fashion for Study 2, although they were not preregistered. The consistency analysis was not preregistered, but the prediction and idea were aligned with our *a priori* theoretical goals and predictions. It was designed and run after we became more familiar with ways to interrogate network structure using neighboring values. The

<sup>1</sup> Motion parameters for one run for one subject were removed due to high collinearity. Motion for the run was minimal, with a mean FD of .0289 mm and no scrubbed volumes.



mPFC prediction for the RSA was accidentally omitted from the preregistration, but was previously identified in the pilot and replicated here.

## Behavioral Results

**Network Centrality Predicts Favorability of Self-Evaluations.** We first sought to replicate the effects of network centrality on self-evaluations demonstrated in Study 2. We tested the interaction of outdegree centrality and valence on self-evaluations, and consistent with our preregistered hypothesis, we found that greater outdegree centrality was associated with more self-descriptiveness for positive traits and less self-descriptiveness for negative traits (see Table 1 for full model results; see Figure 2C for predicted effects). The interaction was decomposed into simple effects, which confirmed that outdegree centrality was negatively associated with self-evaluations for negative traits,  $\beta = -.066$ ,  $SE = .012$ ,  $CI [-.089, -.042]$ ,  $t(6,322) = -5.531$ ,  $p < .001$ ,  $sr^2 = .005$ , and positively associated with self-evaluations for positive traits,  $\beta = .222$ ,  $SE = .011$ ,  $CI [-.061, -.016]$ ,  $t(6,460) = 19.465$ ,  $p < .001$ ,  $sr^2 = .054$ ; see Supplemental Table 14 for model comparisons and parameters for each model; see Supplemental Table 15 for favorability model with covariates.

**Network Relationships Predict Consistency of Self-Evaluations.** We also replicated the pairwise network similarity effects on self-evaluation distance, such that more similar traits (in terms of common neighbors) were evaluated more similarly. The magnitude of these effects was again stronger for positive,  $\beta = -.066$ ,  $SE = .001$ ,  $CI [-.068, -.063]$ ,  $t(467,500) = -46.581$ ,  $p < .001$ , than negative traits,  $\beta = -.043$ ,  $SE = .001$ ,  $CI [-.046, -.040]$ ,  $t(448,800) = -30.033$ ,  $p < .001$ ; see Table 2 for full model results; see Figure 2D for predicted effects; see Supplemental Table 16 for model statistics. Moreover, we replicated the effect of outdegree centrality on network inconsistency, such that outdegree centrality predicted greater consistency of self-evaluations with trait neighbors, supporting the importance of outdegree centrality to maintaining coherence (see Table 3 for full model results; see Supplemental Table 17 for model comparisons and parameters for each model; see Supplemental Table 18 for consistency model with covariates).

**Individual Differences Moderate Centrality Effects on Self-Evaluations.** We further sought to replicate the interactions of self-esteem and depressive symptoms with valence and centrality in predicting favorable self-evaluations demonstrated in Study 2. While the three-way interactions did not replicate in this smaller fMRI sample, the simple effects replicated. Individuals higher in self-esteem,  $\beta = .029$ ,  $SE = .011$ ,  $CI [.006, .051]$ ,  $t(6,458) = 2.492$ ,  $p = .013$ ,  $sr^2 = .001$ , and lower in depressive symptoms,  $\beta = -.034$ ,  $SE = .012$ ,  $CI [-.056, -.011]$ ,  $t(6,312) = -2.915$ ,  $p = .004$ ,  $sr^2 = .001$ , evaluated positive traits higher in outdegree centrality as more self-descriptive and positive traits lower in outdegree centrality as less self-descriptive. Again, self-esteem,  $\beta = .017$ ,  $SE = .012$ ,  $CI [-.016, .040]$ ,  $t(6,320) = 1.448$ ,  $p = .148$ , and depressive symptoms,  $\beta = -.017$ ,  $SE = .012$ ,  $CI [-.041, .006]$ ,  $t(6,176) = -1.425$ ,  $p = .154$ , were not significant in interacting with outdegree centrality for negative traits. Given the replication of the simple effects, we suspect that the lack of replication of the interactive effect is due to the smaller fMRI sample that is underpowered to detect a three-

way cross-level interaction (see Table 4 for full self-esteem model results; see Table 5 for full depressive symptoms model results; see Supplemental Material for self-esteem and depressive symptoms results with covariates).

Additionally, self-esteem and depressive symptoms interacted with outdegree centrality to predict network inconsistency in Study 3, such that outdegree centrality was associated with greater network consistency for individuals higher in self-esteem and lower in depressive symptoms (see Table 6 for full self-esteem moderated consistency model results; see Table 7 for full depressive symptoms moderated consistency model results; see Supplemental Material for self-esteem and depressive symptoms moderated consistency results with covariates).

We measured the associations of other individual differences as well (see Supplemental Table 19 for all individual differences moderated favorability effects and Supplemental Table 20 for all individual differences moderated consistency effects).

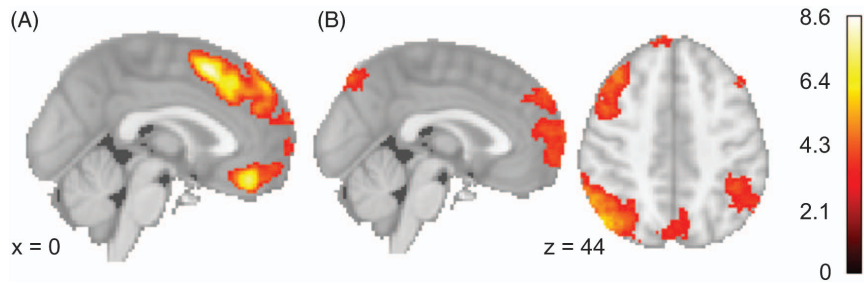
## Neuroimaging Results

### Replication of Self-Relevance Effects

We first sought to replicate the well-documented self-referential effects observed in cortical midline structures. We examined how self-descriptiveness ratings are associated with neural activity. The results supported prior findings on self-referential processing, revealing robust activation in cortical midline structures spanning from precuneus to vmPFC (see Supplemental Figure 7 for multislice mosaic view of self-evaluation effect; Supplemental Table 21 self-evaluation main effect and Supplemental Table 22 for positive self-evaluation > negative self-evaluation effect).

### Effects of Network-Derived Centrality on Brain Activity

Next, we isolated brain activity that parametrically tracks network-defined trait outdegree centrality (see Figure 3A for  $t$ -statistic map at midline slice; see Supplemental Figure 8 for thresholded  $t$ -statistic multislice mosaic map; see Supplemental Figure 9 for unthresholded  $t$ -statistic map). Consistent with our preregistered hypothesis and a previous pilot study (Davis et al., 2014), vmPFC was negatively associated with outdegree centrality ( $-2, 40, -18$ ;  $t = 8.01$ ,  $k = 607$ ; see Table 8 for all clusters). vmPFC may be more involved in processing traits for which there is less available downstream self-knowledge. Conversely, when a trait has more perceived dependencies, it may require less vmPFC recruitment. To account for alternative explanations, we additionally conducted this analysis while controlling for informativeness and social desirability (see Supplemental Figure 10 for midline slice of whole-brain outdegree effect with covariates; see Supplemental Table 23 for all thresholded clusters associated with outdegree, while including covariates) and find that vmPFC is robust even when modeling these additional covariates ( $-2, 38, -18$ ;  $t = 7.09$ ,  $k = 308$ ). We also tested for the conjunction or overlap between the self-descriptiveness and outdegree centrality maps (see Supplemental Figure 11 for conjunction image depicting areas of overlap and nonoverlap between outdegree centrality and self-evaluation). In addition to vmPFC, dorsal mPFC, left angular gyrus, left middle temporal gyrus, bilateral orbitofrontal cortex, and right anterior insula were negatively associated with outdegree centrality.

**Figure 3***Brain Regions Associated With Outdegree Centrality and Trait Similarity*

*Note.* *T*-statistic map. Nonparametric cluster corrected at critical value of  $df = 45$  and  $\alpha = 0.001$ . Bar displays *t*-statistic according to color. (a) Negative association for outdegree effect in vmPFC, dmPFC, and other regions. Figure depicts brain regions that increase in activity when processing less outdegree central traits, and that decrease in activity when processing more outdegree central traits. Traits with more downstream dependents require less vmPFC processing. For example, “thoughtful” is an outdegree central trait which may require less processing by vmPFC during self-reflection. (b) Representational similarity analysis searchlight procedure testing association between network-defined pairwise similarity between traits and neural voxelwise similarity, revealing an amPFC cluster. Figure depicts brain regions for which similarity between voxelwise activation patterns is significantly associated with network similarity between traits. For example, “outgoing” and “sociable” are two similar traits which may share similar patterns of activation in regions such as amPFC. vmPFC = ventral medial prefrontal cortex; amPFC = anterior medial prefrontal cortex; dmPFC = dorsal medial prefrontal cortex.

Only the precuneus positively tracked with outdegree centrality ( $-12, -68, 32$ ;  $t = 6.76$ ,  $k = 405$ ).

We tested the separate effects of outdegree centrality for positive traits (see Supplemental Table 24) and negative traits (see Supplemental Table 25). The contrast comparing outdegree for positive outdegree centrality versus negative outdegree centrality revealed that these effects were stronger for positive outdegree than negative outdegree (see Supplemental Figure 12 and Supplemental Table 26 for positive outdegree centrality > negative outdegree centrality effect). Specifically, outdegree centrality was significantly associated with vmPFC for positive traits and not negative traits, but inspection

of unthresholded, *t*-statistic maps reveals that there is a consistent pattern of negative association with outdegree in the vmPFC region for negative traits, despite not passing our correction and thresholding procedures (see Supplemental Figure 13 for thresholded and unthresholded association with outdegree centrality for positive traits; see Supplemental Figure 14 for thresholded and unthresholded association with outdegree centrality for negative traits).

Indegree centrality, which measures the number traits a given trait is dependent on instead of the number of perceived dependencies, exhibited a distinct pattern of results, primarily being associated with anterior mPFC (see Supplemental Figure 15 and Supplemental Text).

**Table 8**

*Nonparametric Cluster Corrected Maps Indicate Clusters That Are Significantly Associated With Outdegree Centrality, Regardless of Valence*

Parametric regressor of outdegree centrality						
Cluster no.	Positive association	Peak coordinates			<i>T</i>	Size
	Region	<i>x</i>	<i>y</i>	<i>z</i>		
1	Precuneus	-12	-68	32	6.76	405
Cluster no.	Negative association	Peak coordinates			<i>T</i>	Size
	Region	<i>x</i>	<i>y</i>	<i>z</i>		
1	Middle frontal gyrus	-50	14	32	8.55	8,643
	Paracingulate gyrus	-2	14	52	8.44	
	Left orbitofrontal cortex	-48	48	-4	7.97	
2	Left posterior middle temporal gyrus	-60	-40	-12	6.95	2,939
	Left posterior superior temporal gyrus	-54	-42	4	6.79	
3	Right inferior frontal gyrus	56	20	30	6.15	909
4	Left angular gyrus	-52	-70	32	6.03	665
5	Ventromedial prefrontal cortex	-2	40	-18	8.01	607
6	Right lateral orbitofrontal cortex	36	38	-12	7.05	273
7	Right anterior insula	36	26	0	5.65	186

These findings suggest that brain processing during self-reflection is sensitive to the perceived dependency relations described by our trait dependency network, consistent with the hypothesis that people's beliefs about dependency relationships contribute to the organization of self-concept representations in the brain, specifically in vmPFC.

### Network Dependency Relations Predict Similar Activation Patterns

Prior analyses examined how higher-order aspects of network structure, such as outdegree centrality, correlate with the overall magnitude of brain activation during self-reflection. We next sought to interrogate more fine-grained trait-by-trait interrelationships to address where neural activation patterns may differentiate the structural relations between traits in the network. To do so, we tested our prediction that network-defined pairwise trait similarity (i.e., inverse log-weighted similarity) would relate to similarity in neural activation patterns in self- and semantic-processing brain regions (Figure 3b). The whole-brain RSA searchlight procedure revealed large clusters in the anterior medial prefrontal cortex (amPFC), bilateral angular gyrus, bilateral posterior parietal cortex, bilateral rostral lateral prefrontal cortex, occipital pole, right middle temporal gyrus, and right middle frontal gyrus (see Table 9). This suggests that brain regions involved in social and semantic processing represent words similarly based on network similarity. Therefore, given that trait similarity elicits similar voxelwise brain activity, information processed in these regions may represent people's normative beliefs about how traits relate to one another during self-reflection. Further linking how trait similarity is reflected in brain and behavior, trait similarity defined by shared network connections relates to both similarity in self-evaluations and neural representation during self-reflection.

Notably, the RSA for positive traits and negative traits revealed different clusters. Specifically, the RSA for positive traits revealed various clusters consistent with the combined RSA, including the amPFC, whereas the RSA for negative traits revealed a cluster in left posterior parietal cortex ( $-46, -68, 36; t = 4.34, k = 400$ ; See Supplemental Table 27 for the positive RSA and Supplemental

Table 28 for the negative RSA). Given the absence of a significant cluster in amPFC for negative traits, we again examined the unthresholded maps and identified that amPFC exhibits a consistent pattern of association between trait similarity and neural similarity for negative traits, but that it does not pass corrections and thresholding (see Supplemental Figure 16 for unthresholded maps for positive and negative RSAs). Again, we observe valence-based asymmetries in how people process the perceived trait dependencies of traits during self-reflection. Specifically, mPFC may play a role in representing the perceived dependency structure of trait information, but do so to a larger extent for positive more than negative traits.

### Discussion

In Study 3, we replicated the behavioral findings identified in Study 2, including the finding that outdegree centrality predicts more favorable and consistent self-evaluations. We extended these findings by interrogating the neural correlates of outdegree centrality, and regions in which neural similarity is associated with pairwise network-derived trait similarity. We identified an association between outdegree centrality and vmPFC activation, and found that neural similarity is associated with pairwise trait similarity in a more anterior region of mPFC. These results extend prior research on mPFC involvement in self-referential processing, trait knowledge, and conceptual representation by suggesting that this region may play an important role in tracking beliefs about dependency relationships, which may support a positive and coherent self-concept more generally.

In addition to the mPFC, we observed activation tracking our network measures in regions implicated in semantics, concepts, and schemas (Binney & Ramsey, 2020; Gilboa & Marlatte, 2017; Jackson et al., 2015, 2016; Lambon Ralph, 2013; Ralph et al., 2017), such as angular gyrus, middle temporal gyrus, and middle frontal gyrus during self-evaluation. Semantic knowledge allows humans to produce and understand language by transforming sensory inputs into meaningful events. Semantic knowledge relies on conceptual representations, by which humans learn the higher-order relationships among sources of information, and serves to

**Table 9**

*Nonparametric Cluster Corrected Maps Indicate Clusters in Which a Significant Association Was Observed Between Neural Activation Pattern Similarity and Inverse Log-Weighted Similarity as Defined by the Trait Network*

List of representational similarity analysis searchlight regions						
Cluster no.	Region	Peak coordinates			<i>T</i>	Size
		<i>x</i>	<i>y</i>	<i>z</i>		
1	Left posterior parietal cortex	-40	-68	54	6.62	5,102
	Left angular gyrus	-46	-68	38	6.17	
2	Anterior medial prefrontal cortex	-8	62	36	5.99	1,761
3	Left superior frontal gyrus	-44	8	52	5.47	1,392
4	Right posterior parietal cortex	40	-54	54	4.91	1,224
	Right angular gyrus	50	-60	54	4.29	
5	Right middle frontal gyrus	54	40	14	4.45	624
6	Right postcentral gyrus	56	-16	56	4.65	576
7	Left middle temporal gyrus	-66	-50	-8	4.89	460
8	Right rostral lateral prefrontal cortex	-38	58	-2	5.08	349
9	Left rostral lateral prefrontal cortex	40	56	-4	4.20	302
10	Right dorsal lateral prefrontal cortex	28	62	22	4.27	250
11	Occipital pole	22	-99	4	3.99	166

promote knowledge generalization across experiences and contexts. Given that the task requires language comprehension and inferences about semantic relationships, it is likely that these regions may be performing a similar function with this task as in other semantic and concept related research. Interestingly, we did not find anterior temporal lobe associations with either outdegree centrality or pairwise similarity despite the fact that damage in this region tends to be associated with semantic deficits in a number of patient groups (Bonner & Price, 2013). However, this region is not always found to be activated in functional imaging studies, and its absence here may simply mean that it activates similarly across trait words as opposed to not being involved at all. The involvement of the angular gyrus is also theoretically interesting, as it has been implicated as a cross-modal integrative hub that operates in tandem with other systems to comprehend and give meaning to experiences (Seghier, 2013; Wagner et al., 2015), and in binding relational information (Shimamura, 2011; Treisman & Gelade, 1980) or recombining semantic information (Price et al., 2016; Yazar et al., 2014). The angular gyrus may support inferential processing in vmPFC to incorporate novel dependent traits, serving as an integrative hub during self-referential processing. We also found that voxelwise activation patterns in posterior parietal cortex, a region implicated in structure learning (Summerfield et al., 2020), were associated with network-defined trait similarity. This region may support relational inferences in nonspatial domains, including relational inferences within the self-concept as well. Altogether, regions involved in semantics and structure learning may support the organization of normative latent trait structure that contributes to self-concept structure.

## General Discussion

We present a new model of self-concept representation that unites neurobiological and psychological research on the self into a common computational framework. We examine the self as a system of interdependent trait self-perceptions by developing a network model of perceived dependency relations between traits from human raters (Study 1). We then tested predictions from this trait dependency network model for both behavioral self-evaluations and brain activation in two independent samples of participants (Study 2 and Study 3). We found that the number of perceived trait dependencies (i.e., outdegree centrality) was associated with more favorable and consistent self-evaluations. Further, individuals with higher self-esteem and fewer depressive symptoms differentiated between positive traits with higher and lower outdegree more in their self-evaluations and self-evaluated more consistently between higher outdegree traits and their neighbors. Together, these results are consistent with our hypothesis that traits with more perceived dependencies are critical for maintaining self-concept positivity and coherence, supporting our trait network model as a viable explanation of the cognitive mechanisms contributing to self-concept representation. In Study 3, we extended this model to an fMRI experiment and found vmPFC activation tracked outdegree centrality during self-evaluations, and activation patterns in amPFC reflected the perceived dependency relationships described by the network. We provide a formal model of self-representation based on people's perceptions of trait dependencies that connects disparate findings from the psychological and neuroscience literatures on semantics, concepts, and the self.

## vmPFC's Potential Role in Self-Concept Organization and Inference

Our approach extends previous research on the vmPFC's role in self-reflection. We focus on the vmPFC specifically given our *a priori* interest in the region due to its prevalence in self-referential (D'Argembeau, 2013) and social cognitive (Mitchell, 2009) processes. We found that vmPFC activation decreases to the extent that a trait is higher in outdegree centrality, such that vmPFC function may process beliefs about trait dependencies during trait self-evaluations. We note that the current finding replicates prior pilot unpublished analyses on an independent data set (Davis et al., 2014). We suggest that outdegree centrality (or more generally, a self-belief with many implications) may be particularly important to the maintenance of coherence. This is because traits with greater outdegree centrality enable more opportunities for people to generate contradictions with downstream traits that depend on them if, for example, a person endorses a downstream trait but not the upstream trait that people believe it depends on. In this way, coherence is a global property of having noncontradictory beliefs about the self, which can be preserved, in part, by focusing on traits that are more central and have many dependencies (i.e., higher outdegree traits). Given that vmPFC activity is sensitive to differences in the outdegree centrality of traits, this suggests a potential role for vmPFC function in maintaining self-concept coherence.

Self-schema theory conceives of the self-concept as a knowledge structure organized around centrally defined features (Markus, 1977), and extending this, we find that the vmPFC—a region involved in schema-processing (Gilboa & Marlatte, 2017)—also processes the normative centrality of self-relevant features. By relating vmPFC activation during self-reflection to formally defined measures of perceived trait structure, we build stronger connections between our work and other work on vmPFC's contributions to general social and nonsocial conceptual representations (for a recent review, see Zeithamova et al., 2019). While the vmPFC is involved in myriad different processes (Koban et al., 2021), the regions that coactivate with it can help contextualize what function it may be involved in for a given mental process. During self-processing, vmPFC often coactivates with precuneus and posterior cingulate (Denny et al., 2012; Northoff et al., 2006), as observed in our self-evaluation effect. The patterns of brain activity associated with outdegree centrality are consistent with vmPFC's involvement in schematic knowledge (Gilboa & Marlatte, 2017), concept representation (Zeithamova et al., 2019), and semantic memory more broadly (Binder et al., 2009). For instance, the vmPFC is often involved in detecting the congruence of information with existing schematic knowledge during encoding and retrieval (e.g., incongruent: "polar bear" and "desert"; congruent: "stapler" and "office"), reflecting its potential role in schema generalization (Gilboa & Marlatte, 2017; Spalding et al., 2015). Consistent with this, in the current research, vmPFC recruitment may reflect generalization from a trait self-evaluation to a structure of dependent traits forming a self-concept. Lower outdegree centrality traits may require greater inferential processing to detect potential traits that may be implicated by the current self-evaluation, as people believe fewer traits depend on them. When generalizing a self-evaluation across fewer possible implications, the scarcity of downstream information may require greater involvement of vmPFC.

An alternative possibility is that it is not outdegree centrality per se that is driving vmPFC activation, but rather another property



correlated with outdegree centrality in the present sample. We eliminate a number of possible alternative explanations by showing that the effects of outdegree remain when controlling for a variety of other trait characteristics, such as desirability, breadth, and informativeness. However, given that representations of traits are complex and multifaceted, it will be critical for future work to continue to build additional realism and complexity into our trait dependency network model to understand how a wider range of normative characteristics may contribute to, or follow from, perceptions of trait interrelatedness. While we acknowledge there are likely other factors that play a role, our findings show that people's beliefs about directed relations among trait self-beliefs relate to how people maintain positive and coherent self-views. Therefore, while we cannot rule out all alternatives, we can predict a variety of effects from people's beliefs about dependency that are replicable and distinct from a number of normative trait characteristics.

Our research also advances our understanding of how the brain encodes and organizes multifaceted and interrelated self-relevant information. While recent research has begun to examine psychological and neural representations of trait knowledge structure during person perception (Stolier et al., 2020; Tamir & Thornton, 2018), the present study extends this burgeoning area by showing how the brain also represents interrelationships between traits during self-representation. By analyzing how distributed patterns of activity relate to network-defined similarity between traits, we move beyond considering mere involvement of brain regions in tracking higher-order structured relationships between self-relevant traits (i.e., outdegree centrality), and instead consider the content of the network's representation. We identify regions associated with social- (e.g., amPFC) and semantic-processing (e.g., inferior frontal gyrus, middle temporal gyrus) as distinctly involved in representing trait structure during self-reflection. Therefore, the normative trait dependency network may serve as a useful explanation for how the brain processes multifaceted and interrelated self-relevant information into a coherent self-concept. As the network was developed via the nomination of perceived dependency relationships, we propose that structured interrelationships between traits may be *encoded* in semantic-processing brain regions (Jackson et al., 2016; Ralph et al., 2017). In tandem, the amPFC may *integrate* these structured semantic interrelationships and organize them into a higher-order self-representation. While prior work describes the mPFC's role in spontaneous trait inferences (Ma et al., 2014) and in representing personality and trait dimensions in person perception (Hassabis et al., 2014; Thornton & Mitchell, 2018), we extend these findings by formalizing a structure of trait relationships that allows us to interrogate different features of these relationships that may be important for psychological and neurobiological self-processing. Collectively, these findings provide a framework for understanding the relationship between the structure of trait beliefs, the self-concept, and the content of its neural representation.

Our approach for studying the effects of perceived trait dependencies on self-evaluations bears important implications on computational language analysis and semantics. Word meanings are often difficult to find universal agreement upon, and it is difficult to find necessary or sufficient criteria for defining a concept. As such, linguistic work on semantics has generally relied on corpus analysis to uncover word meanings based on whether words with similar meanings co-occur with one another. However, this approach can be difficult to apply to adjectives, such as personality traits, as they may

not co-occur with similar adjectives and the meaning of traits may be highly contextual. Moreover, even if co-occurrences were a useful tool for understanding the meaning behind personality traits, they would not provide insight into coherence. Given that coherence reflects the desire or state of maintaining noncontradictory beliefs, which reflects consistency between a belief and its dependent beliefs, a model of coherence requires the incorporation of directionality into how we understand the meaning of personality traits. Here, we find that simply asking participants what they perceive to be true of words provides insight into how coherence is accomplished. This surveying of participants on normative beliefs and attitudes is common practice in social and personality psychology (e.g., our assessment of the normative desirability and breadth of traits), and we must rely on consensus among participants for developing our network of trait dependencies, rather than relying on our intuitions to characterize pairwise dependencies among traits. When directionality matters and we do not trust our intuitions as researchers to ad hoc determine the bins that stimuli belong to, we must trust the data as a basis for understanding the directions and dependencies among traits. Indeed, we find that this consensus among participants in terms of pairwise trait dependencies is robust in terms of the validity, reliability, and stability of the measures. Importantly, other theories that only consider single, nondyadic relationships amongst trait words do not anticipate the findings we reveal using our directed network approach.

### Psychological Differences Across Valence, Mental Health, and Culture

Our trait network approach sheds light on how self-concept positivity may be linked with self-concept coherence. Decades of research suggest that individuals are motivated to hold both positive (Taylor & Brown, 1988) and coherent (i.e., noncontradictory; Swann et al., 2003) self-views. One possibility is that people may simultaneously accomplish these dual motives by evaluating most favorably and consistently on traits with more perceived dependencies. By evaluating higher outdegree traits more favorably, people may be able to also evaluate favorably on their downstream dependent traits without contradiction, thus achieving coherence amongst one's self-views and beliefs about how traits depend upon one another. By permitting favorable evaluations on downstream dependents, evaluating oneself favorably on higher outdegree traits may thus simultaneously maximize coherence and propagate positivity. While prior psychological (Alicke & Sedikides, 2009; Dunning, 1999; Swann et al., 2003) and neuroscientific research (Beer, 2007; Hughes & Zaki, 2015; Sharot & Garrett, 2016) has described motivations for either positivity or coherence, our findings extend these insights by characterizing how self-concept positivity and coherence may mutually depend on people's beliefs about trait dependency relations.

Interestingly, behavioral effects of network measures on self-evaluations were generally stronger for positive traits than negative traits, an effect that was also mirrored in their brain activation. One possibility is that, given that positive information is generally perceived as more similar than negative (Alves et al., 2017a, 2017b; Koch et al., 2016; Unkelbach et al., 2020), evaluations of negative traits will generalize less to other negative traits because they are considered more distinct, whereas evaluations of positive traits will generalize more to other positive traits (Gräf & Unkelbach,

2016, 2018). This may enable people to more readily perceive relationships among positive versus negative traits and generalize across them more effectively. Alternatively, people may be motivated to discount negative information and its implications for the self (Baumeister et al., 2001; Taylor, 1991), and instead engage in fine-grained processing of positive information as it relates to the self (Alicke & Sedikides, 2009; Sedikides & Gregg, 2008). Future research should examine whether valence asymmetries in how trait interdependencies predict self-views are better explained by a motivational account whereby people discount negative information, a representational account whereby negative information is structured more sparsely and distinctly than positive information, or a combination of both.

If people maintain self-concept positivity and coherence by being tuned to how self- and trait knowledge is structured, abnormalities in how people process this structure may relate to mental health symptomatology. Prior research has described aggregate reductions in positive self-views (Bernet et al., 1993; Tarlow & Haaga, 1996) and a preference for self-consistent negative information (North & Swann, 2009; Swann et al., 1992) as a function of depressive symptoms and low self-esteem. The current findings extend these observations and offer potential insight into one possibility for how reductions in self-concept positivity and coherence may emerge among individuals with depressive symptoms and low self-esteem. We show that those higher in self-esteem and lower in depression may not only evaluate more positively about themselves overall, but may also evaluate most positively and consistently on traits with more perceived dependencies. It is possible that a contributor to how psychologically well-adjusted individuals maintain positive and consistent self-views is by weighting trait dependencies during self-evaluations. By maximizing positivity and consistency among traits with the most perceived dependencies, individuals higher in self-esteem and lower in depression may thereby maintain a positive and coherent global self-concept.

Our findings raise the possibility that self-concept positivity and coherence partially depend on how individuals process structured trait relationships when they self-reflect. Given that the outdegree connections in the trait dependency network were formed from independent ratings of trait dependencies, there is no *a priori* semantic reason outside of the current theory to suggest that traits with higher outdegree centrality would be less related to self-evaluations for individuals higher in depressive symptoms or lower in self-esteem. This helps to constrain the possible explanations for how the positivity or negativity of self-views may relate to trait structure. While the directionality of the effect is unclear, we suggest that the processing of trait dependencies may help maintain a more positive and coherent self-concept, by propagating positivity and maintaining noncontradictory self-views on traits with more perceived consequences.

An open question is whether greater depressive symptoms or lower self-esteem *leads* to failures to propagate positivity and maintain coherence, or whether they *arise* from failures to propagate positivity and maintain coherence. While past work suggests that depressive symptoms are associated with more tightly organized schemas (Malle & Horowitz, 1995), cognitive rigidity, and inflexibly negative self-views (Meiran et al., 2011; Stange et al., 2017), we highlight an important distinction between rigidity and coherence. Inflexibly negative self-views can threaten coherence if a person fails to maintain consistency between traits and their perceived

dependents, or fails to flexibly update traits due to the rigidity of self-views. An alternative interpretation is that individuals higher in depressive symptoms exhibit the same sensitivity to central, coherence-preserving traits as those lower in depressive symptoms, but rather have an organization of perceived dependencies that differs from the existing network (e.g., depressed individuals evaluate favorably on central traits, but disagree that “thoughtful” is a central trait). To investigate this, separate dependency networks could be created for depressed and nondepressed groups to test whether the resulting dependency networks are substantially different from one another.

Given the evidence of cultural differences in self-enhancement biases (Heine et al., 1999; Heine & Lehman, 1995) and in neurobiological self-referential function (Hampton & Varnum, 2018; Han & Northoff, 2008; Kitayama & Park, 2014), there may be cultural differences in the effect of outdegree centrality on favorability of self-judgments and vmPFC function. However, it is important to consider whether these possible cultural differences emerge due to differences in nodal weights applied to a trait network (i.e., self-evaluations) or differences in connections in the trait dependency network structure itself (i.e., normative dependency beliefs). One possibility is that cultures differ in the weights they apply to traits based on their network positions and features. For example, individuals from one culture may on average evaluate higher on “conscientious” while individuals from a different culture may evaluate lower on “conscientious.” In this scenario, individuals from both cultures may agree on the relations between traits, such that individuals from both cultures normatively believe that “conscientious” has similar dependent traits. The current results would thus be relatively culture-independent as the structure of the network will generalize to other cultures. Alternatively, cultures may differ in both the weights they apply to traits in the network and in their normative beliefs about dependencies. For example, individuals in one culture may believe that many traits depend on “conscientious” whereas individuals from a different culture may believe that few traits do. In this case, the trait networks would be culture-dependent, and separate networks would need to be constructed for different cultures. Future research should aim to collect and compare a variety of trait networks normed to different culture-specific semantic beliefs to answer these key questions about the cross-cultural generalizability of this work.

The configuration of the self-concept may also diverge cross-culturally as a function of whether individuals within a culture construe the self as independent or interdependent (Markus & Kitayama, 1991). These self-construals can influence the extent to which people value self-specific versus other-specific attributes as important to their self-concept, thereby facilitating information processing for attributes and behaviors that are consistent with these more elaborated schemas (Markus, 1977). From this perspective, cultural differences in self-construal could be explained by considering the self-concept as a hierarchical structure (McConnell, 2011; Schell et al., 1996) with superordinate attributes (e.g., “independent”) that organize subordinate features. For example, in American culture, the trait “independent” may be superordinate to other traits that people of that culture tend to value. A superordinate trait in a hierarchical structure may parallel higher outdegree centrality traits in that they are similarly important to the coherence of the structure due to having many dependents. However, superordinate categories (e.g., “animal”) are often less informative than lower

level “basic-level” categories (e.g., “bird”) about what other features or traits a member of that category may possess (Cortner & Gluck, 1992). Applying this observation to personality traits, in American culture, the trait “independent” might be less informative about individuals within the culture (as opposed to taxonomic differences between cultures) because many people possess it.

## Future Directions

The knowledge structures that contribute to the self-concept, and what aspects are characterized by a hierarchical or the flat network structure of our current model, remains an open question. This is partially because there have not been many direct attempts to develop a computational model that represents hierarchies amongst large numbers of traits to compare to our network model here. However, there is some reason to expect, given our results, that a model with a hierarchical structure would not capture participants’ representations of trait dependency. For example, although our network has directed connections between traits, we also find reciprocal and cyclical connections between traits, suggesting that, for the traits that we observed, there is not a clear hierarchy where any trait provides a superordinate organization that all other traits must follow. Future research will need to carefully test these two ways of understanding conceptual organization. Our current approach is useful in this regard as it provides one example of how to formally articulate, with a network model, how self-knowledge may be structured within a self-concept, and provides a platform for future research. For example, one possible formulation of the self-concept to explore in future research may be a multilevel network that contains both hierarchical as well as flat/non-hierarchical inter-related components, by which a network of interrelated traits is superordinate to bodily and mental states, and subordinate to group memberships and social identities (e.g., Tamir & Thornton, 2018 for a related approach applied to person perception).

In the present research, we develop a highly expressive and generative computational framework of trait dependencies that informs neuroscience and psychology perspectives on self- and social-representation. We focus on how normative beliefs about trait dependencies lead to positivity and coherence across individuals in representations of the self-concept. However, given that normative trait beliefs likely affect attributions in a variety of domains, this approach could be generative for research programs seeking to understand how conceptual representations influence social judgments more generally, such as how people maintain coherent impressions of other people.

Future research may consider what implications the current findings have on the representations of other self-knowledge, beyond traits. For example, although this network is defined by normative beliefs about dependency relations that characterize how people generally represent trait relationships, this framework may provide insight into how people represent more idiosyncratic and personalized self-concept structures that include autobiographical memories, personal narratives, and social identities. It may be that people also reflect more favorably and consistently upon more central social identities or core autobiographical memories upon which other identities or memories depend. Future research may further interrogate commonalities and differences in nomothetic and idiographic components of self-processing and self-structure.

Future research should also explore the extent to which consistent and favorable self-evaluations on higher outdegree traits may contribute to person-level self-concept coherence and stability. As we discuss above, coherence is an emergent property of the network characterized by noncontradictory self-beliefs, and thus cannot be measured at the trait level. Future research may benefit from further linking self-concept coherence at the person level to self-evaluations within a trait dependency network. For example, individuals with borderline personality disorder (BPD) often report being uncertain of their own personality traits and exhibit highly malleable, situation-based identities (Flury & Ickes, 2007). These BPD individuals are known to have a weaker “sense of self” and exhibit a lack of understanding of oneself along with sudden shifts in feelings, opinions, and values (Briere & Runtz, 2002; Gunderson, 2009; Kreisman & Straus, 2021). Identifying whether individuals known to have greater instability within the self-concept also self-evaluate less consistently and favorably on higher outdegree centrality trait and exhibit differential processing of outdegree centrality in vmPFC—akin to deficits in processing schema-congruence observed among vmPFC lesion patients (Ghosh & Gilboa, 2014; Spalding et al., 2015)—may further reinforce the present theory.

## Limitations

In addition to the contributions and future extensions of the present research, it does carry some limitations. While our selection of traits was extensive, findings may depend partially on the sampling of traits and participants used to construct network. Our neuroimaging findings drew on a sample of English-speaking university student participants, and thus findings may differ among different language speakers. However, our findings did generalize across samples with quite different racial demographics, reflecting generalizability across a variety of demographic groups. Additionally, given that the model developed here was a normative trait dependency network constructed via an independent set of raters, this approach cannot speak directly to how people idiosyncratically structure self-specific beliefs. The model is nomothetic, in the sense that it describes general beliefs about trait dependency, shared among an independent group of raters, but the same logic of dependency relations could also be applied to more idiographic self-concept representations. However, assessing idiosyncrasies in dependency beliefs was not the goal of the present research: We do not claim that the network or outdegree centrality is self-specific, as it is a model of people’s beliefs about traits dependencies. Rather we highlight its importance to self-referential processing and that individuals’ weighting of traits and their directed connections may contribute to self-concept positivity and coherence.

## Conclusion

The contents of the self-concept must necessarily consist of both structured idiosyncratic (i.e., episodic) and normative (i.e., semantic) knowledge (e.g., Kihlstrom et al., 2003). Social psychology has traditionally focused on the self-concept as an idiosyncratic, idiographic structure, whereas cognitive psychology has assumed a “ground truth” to generalized, nomothetic cognitive conceptual models. Here, we extend prior theoretical frameworks of the self-concept as an idiosyncratic and idiographic self-belief structure (Markus & Wurf, 1987; McConnell, 2011), by developing a



nomothetic, normative trait dependency network that formalizes how people construct inferences about the self by idiosyncratically assigning weights (i.e., self-evaluations) to traits based on their normatively perceived relationships to other traits (e.g., outdegree, similarity). Our results suggest that people maintain self-concept coherence and positivity by monitoring the downstream consequences of their self-evaluations to avoid contradictory self-evaluations between traits and their dependents and to propagate positivity. On the implementation level, mPFC—a region commonly implicated in self-reflection, schema inferences, and concept organization—represents the normative trait beliefs that contribute to self-concept structure.

While self-concept coherence has typically been defined as being achieved by aligning experiences with self-views (e.g., Swann et al., 2003), we add to this perspective by illustrating how the drive to maintain consistency, by avoiding contradictory self-views between traits and their dependents, can simultaneously maximize coherence and positivity. In this way, the drive to maintain self-concept coherence may also stem from representations of dependency relations amongst traits. In other words, self-concept coherence is not merely achieved by seeking consistent experiences, but is also an emergent property of the concept structure itself. One possibility for how people achieve local self-consistency and global coherence is that, rather than maintaining a global representation of the self that may be computationally costly, people may instead construct inferences of themselves “on the fly” by detecting the relationships between self-perceptions. The self-concept may be dynamically informed by the relatedness of a recent self-perception to other proximal, possible self-perceptions, perhaps providing a formalized approach for understanding the underpinnings of the contextualized and situationally determined working self-concept (Markus & Kunda, 1986). Given the generative nature of the current perspective, it may shed light on the stability and malleability of dynamic self-related inferences (Elder et al., 2022), and how self-concept structure relates to mental health or psychological well-being.

### Citation Diversity Statement

Recent work in several fields of science has identified a bias in citation practices such that articles from women and other minority scholars are under-cited relative to the number of such articles in the field (Caplar et al., 2017; Dion et al., 2018; Dworkin et al., 2020; Maliniak et al., 2013; Mitchell et al., 2013). Here, we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. First, we obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman (Dworkin et al., 2020; Zhou et al., 2020). By this measure (and excluding self-citations to the first and last authors of our current article), our references contain 8.88% woman(first)/woman(last), 11.38% man/woman, 22.01% woman/man, and 57.74% man/man. This method is limited in that a) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity and b) it cannot account for intersex, nonbinary, or transgender people. Second, we obtained predicted racial/ethnic category of the first and last author of each reference by databases that store the probability of a first and last name being carried by an author of color (Ambekar et al., 2009; Sood & Laohaprapanon,

2018). By this measure (and excluding self-citations), our references contain 4.27% author of color (first)/author of color(last), 12.66% White author/author of color, 15.98% author of color/White author, and 67.08% White author/White author. This method is limited in that (a) names and Florida voter data to make the predictions may not be indicative of racial/ethnic identity, and (b) it cannot account for Indigenous and mixed-race authors, or those who may face differential biases due to the ambiguous racialization or ethnicization of their names. We look forward to future work that could help us to better understand how to support equitable practices in science.

### 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>
- Alloy, L. B., Black, S. K., Young, M. E., Goldstein, K. E., Shapero, B. G., Stange, J. P., Boccia, A. S., Matt, L. M., Boland, E. M., Moore, L. C., & Abramson, L. Y. (2012). Cognitive vulnerabilities and depression versus other psychopathology symptoms and diagnoses in early adolescence. *Journal of Clinical Child and Adolescent Psychology*, 41(5), 539–560. <https://doi.org/10.1080/15374416.2012.703123>
- Alves, H., Koch, A., & Unkelbach, C. (2017a). The “common good” phenomenon: Why similarities are positive and differences are negative. *Journal of Experimental Psychology: General*, 146(4), 512–528. <https://doi.org/10.1037/xge0000276>
- Alves, H., Koch, A., & Unkelbach, C. (2017b). Why good is more alike than bad: Processing implications. *Trends in Cognitive Sciences*, 21(2), 69–79. <https://doi.org/10.1016/j.tics.2016.12.006>
- Ambekar, A., Ward, C., Mohammed, J., Male, S., & Skiena, S. (2009). Name-ethnicity classification from open sources. *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 49–58). Association for Computing Machinery. <https://doi.org/10.1145/1557019.1557032>
- Anderson, J. R., & Bower, G. H. (1973). *Human associative memory* (p. 524). V. H. Winston & Sons.
- Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, 9(3), 272–279. <https://doi.org/10.1037/h0025907>
- Andersson, J. L. R., Jenkinson, M., & Smith, S. (2007). *Non-linear registration aka spatial normalisation* (FMRIB Technical Report TR07JA2). <https://www.scienceopen.com/document?vid=13f3b9a9-6e99-4ae7-bea2-c1bf0af8ca6e>
- Andersson, J. L. R., Skare, S., & Ashburner, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. *NeuroImage*, 20(2), 870–888. [https://doi.org/10.1016/S1053-8119\(03\)00336-7](https://doi.org/10.1016/S1053-8119(03)00336-7)
- Araujo, H. F., Kaplan, J., & Damasio, A. (2013). Cortical midline structures and autobiographical-self processes: An activation-likelihood estimation meta-analysis. *Frontiers in Human Neuroscience*, 7, Article 548. <https://doi.org/10.3389/fnhum.2013.00548>
- 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/10.1016/j.jml.2012.11.001>
- Barton, K. (2009). *MuMIn: multi-model inference*. <http://r-forge.r-project.org/projects/mumin/>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), 323–370. <https://doi.org/10.1037/1089-2680.5.4.323>



- Beck, A. T., Steer, R. A., & Brown, G. K. (1987). *Beck depression inventory*. Harcourt Brace Jovanovich.
- Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in fMRI. *NeuroImage*, 20(2), 1052–1063. [https://doi.org/10.1016/S1053-8119\(03\)00435-X](https://doi.org/10.1016/S1053-8119(03)00435-X)
- Beer, J. S. (2007). The default self: Feeling good or being right? *Trends in Cognitive Sciences*, 11(5), 187–189. <https://doi.org/10.1016/j.tics.2007.02.004>
- Beer, J. S., & Hughes, B. L. (2010). Neural systems of social comparison and the “above-average” effect. *NeuroImage*, 49(3), 2671–2679. <https://doi.org/10.1016/j.neuroimage.2009.10.075>
- Bernet, C. Z., Ingram, R. E., & Johnson, B. R. (1993). Self-esteem. In C. G. Costello (Ed.), *Symptoms of depression* (pp. 141–159). Wiley.
- Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12), 2767–2796. <https://doi.org/10.1093/cercor/bhp055>
- Binney, R. J., & Ramsey, R. (2020). Social semantics: The role of conceptual knowledge and cognitive control in a neurobiological model of the social brain. *Neuroscience and Biobehavioral Reviews*, 112, 28–38. <https://doi.org/10.1016/j.neubiorev.2020.01.030>
- Bonner, M. F., & Price, A. R. (2013). Where is the anterior temporal lobe and what does it do? *The Journal of Neuroscience*, 33(10), 4213–4215. <https://doi.org/10.1523/JNEUROSCI.0041-13.2013>
- Bower, G. H., & Gilligan, S. G. (1979). Remembering information related to one's self. *Journal of Research in Personality*, 13(4), 420–432. [https://doi.org/10.1016/0092-6566\(79\)90005-9](https://doi.org/10.1016/0092-6566(79)90005-9)
- Briere, J., & Runtz, M. (2002). The Inventory of Altered Self-Capacities (IASC): A standardized measure of identity, affect regulation, and relationship disturbance. *Assessment*, 9(3), 230–239. <https://doi.org/10.1177/1073191102009003002>
- Caplar, N., Tacchella, S., & Birrer, S. (2017). Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy*, 1(6), 1–5. <https://doi.org/10.1038/s41550-017-0141>
- Chan, S. C. Y., Niv, Y., & Norman, K. A. (2016). A probability distribution over latent causes, in the orbitofrontal cortex. *The Journal of Neuroscience*, 36(30), 7817–7828. <https://doi.org/10.1523/JNEUROSCI.0659-16.2016>
- Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, 51(5), 2022–2038. <https://doi.org/10.3758/s13428-019-01273-7>
- Chavez, R. S., Heatherton, T. F., & Wagner, D. D. (2017). Neural population decoding reveals the intrinsic positivity of the self. *Cerebral Cortex*, 27(11), 5222–5229. <https://doi.org/10.1093/cercor/bhw302>
- Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292), 1464–1468. <https://doi.org/10.1126/science.aaf0941>
- Cortner, J., & Gluck, M. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, 111(2), 291–303. <https://doi.org/10.1037/0033-2909.111.2.291>
- Craik, F. I. M., Moroz, T. M., Moscovitch, M., Stuss, D. T., Winocur, G., Tulving, E., & Kapur, S. (1999). In search of the self: A positron emission tomography study. *Psychological Science*, 10(1), 26–34. <https://doi.org/10.1111/1467-9280.00102>
- D'Argembeau, A. (2013). On the role of the ventromedial prefrontal cortex in self-processing: The valuation hypothesis. *Frontiers in Human Neuroscience*, 7, Article 372. <https://doi.org/10.3389/fnhum.2013.00372>
- Davidson, D., & LePore, E. (1986). A coherence theory of truth and knowledge. In E. Sosa, J. Kim, J. Fantl, & M. McGrath (Eds.), *Epistemology: An Anthology*, (2nd ed., pp.124–133). Wiley-Blackwell.
- Davis, T., Hughes, B., & Beer, J. (2014, May 20). *Structure of conceptual relations how positivity bias reflects the structure of self concepts*. Association for Psychological Science Convention. <https://doi.org/10.13140/RG.2.2.25347.84003>
- Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, 24(8), 1742–1752. [https://doi.org/10.1162/jocn\\_a\\_00233](https://doi.org/10.1162/jocn_a_00233)
- Descartes, R. (1641/1998). *Meditations and other metaphysical writings*. Penguin. (Original work published 1641)
- Dion, M. L., Sumner, J. L., & Mitchell, S. M. (2018). Gendered citation patterns across political science and social science methodology fields. *Political Analysis*, 26(3), 312–327. <https://doi.org/10.1017/pan.2018.12>
- Dufner, M., Gebauer, J. E., Sedikides, C., & Denissen, J. J. A. (2019). Self-enhancement and psychological adjustment: A meta-analytic review. *Personality and Social Psychology Review*, 23(1), 48–72. <https://doi.org/10.1177/1088868318756467>
- Dunning, D. (1999). A newer look: Motivated social cognition and the schematic representation of social concepts. *Psychological Inquiry*, 10(1), 1–11. [https://doi.org/10.1207/s15327965pli1001\\_1](https://doi.org/10.1207/s15327965pli1001_1)
- Dworkin, J. D., Linn, K. A., Teich, E. G., Zurn, P., Shinohara, R. T., & Bassett, D. S. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, 23(8), 918–926. <https://doi.org/10.1038/s41593-020-0658-y>
- Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., & Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine*, 27(29), 6137–6157. <https://doi.org/10.1002/sim.3429>
- 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>
- Flury, J. M., & Ickes, W. (2007). Having a weak versus strong sense of self: The Sense of Self Scale (SOSS). *Self and Identity*, 6(4), 281–303. <https://doi.org/10.1080/15298860601033208>
- Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, 53, 104–114. <https://doi.org/10.1016/j.neuropsychologia.2013.11.010>
- Giesler, R. B., Josephs, R. A., & Swann, W. B., Jr. (1996). Self-verification in clinical depression: The desire for negative evaluation. *Journal of Abnormal Psychology*, 105(3), 358–368. <https://doi.org/10.1037/0021-843X.105.3.358>
- Gilboa, A., & Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory. *Trends in Cognitive Sciences*, 21(8), 618–631. <https://doi.org/10.1016/j.tics.2017.04.013>
- Gräf, M., & Unkelbach, C. (2016). Halo effects in trait assessment depend on information valence: Why being honest makes you industrious, but lying does not make you lazy. *Personality and Social Psychology Bulletin*, 42(3), 290–310. <https://doi.org/10.1177/0146167215627137>
- Gräf, M., & Unkelbach, C. (2018). Halo effects from agency behaviors and communion behaviors depend on social context: Why technicians benefit more from showing tidiness than nurses do. *European Journal of Social Psychology*, 48(5), 701–717. <https://doi.org/10.1002/ejsp.2353>
- Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 493–498. <https://doi.org/10.1111/2041-210X.12504>
- Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist*, 35(7), 603–618. <https://doi.org/10.1037/0003-066X.35.7.603>
- Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1), 63–72. <https://doi.org/10.1016/j.neuroimage.2009.06.060>
- Gunderson, J. G. (2009). Borderline personality disorder: Ontogeny of a diagnosis. *The American Journal of Psychiatry*, 166(5), 530–539. <https://doi.org/10.1176/appi.ajp.2009.08121825>
- Hadjichristidis, C., Sloman, S., Stevenson, R., & Over, D. (2004). Feature centrality and property induction. *Cognitive Science*, 28(1), 45–74. [https://doi.org/10.1207/s15516709cog2801\\_2](https://doi.org/10.1207/s15516709cog2801_2)

- Hampson, S. E., Goldberg, L. R., & John, O. P. (1987). Category-breadth and social-desirability values for 573 personality terms. *European Journal of Personality*, 1(4), 241–258. <https://doi.org/10.1002/per.2410010405>
- Hampton, R. S., & Varnum, M. E. W. (2018). Do cultures vary in self-enhancement? ERP, behavioral, and self-report evidence. *Social Neuroscience*, 13(5), 566–578. <https://doi.org/10.1080/17470919.2017.1361471>
- Han, S., & Northoff, G. (2008). Culture-sensitive neural substrates of human cognition: A transcultural neuroimaging approach. *Nature Reviews Neuroscience*, 9(8), 646–654. <https://doi.org/10.1038/nrn2456>
- Hanke, M., Halchenko, Y. O., Sederberg, P. B., Olivetti, E., Fründ, I., Rieger, J. W., Herrmann, C. S., Haxby, J. V., Hanson, S. J., & Pollmann, S. (2009). PyMVPA: A unifying approach to the analysis of neuroscientific data. *Frontiers in Neuroinformatics*, 3, Article 3. <https://doi.org/10.3389/neuro.11.003.2009>
- Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, 24(8), 1979–1987. <https://doi.org/10.1093/cercor/bht042>
- Heine, S. J., & Lehman, D. R. (1995). Cultural variation in unrealistic optimism: Does the West feel more vulnerable than the East? *Journal of Personality and Social Psychology*, 68(4), 595–607. <https://doi.org/10.1037/0022-3514.68.4.595>
- Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review*, 106(4), 766–794. <https://doi.org/10.1037/0033-295X.106.4.766>
- Huff, C., & Tingley, D. (2015). “Who are these people?” Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, 2(3), 1–12. <https://doi.org/10.1177/2053168015604648>
- Hughes, B. L., & Beer, J. S. (2013). Protecting the self: The effect of social-evaluative threat on neural representations of self. *Journal of Cognitive Neuroscience*, 25(4), 613–622. [https://doi.org/10.1162/jocn\\_a\\_00343](https://doi.org/10.1162/jocn_a_00343)
- Hughes, B. L., & Zaki, J. (2015). The neuroscience of motivated cognition. *Trends in Cognitive Sciences*, 19(2), 62–64. <https://doi.org/10.1016/j.tics.2014.12.006>
- Hume, D. (1739/2003). *A treatise of human nature*. Courier Corporation. (Original work published 1739)
- Jackson, R. L., Hoffman, P., Pobric, G., & Lambon Ralph, M. A. (2015). The nature and neural correlates of semantic association versus conceptual similarity. *Cerebral Cortex*, 25(11), 4319–4333. <https://doi.org/10.1093/cercor/bhv003>
- Jackson, R. L., Hoffman, P., Pobric, G., & Lambon Ralph, M. A. (2016). The semantic network at work and rest: Differential connectivity of anterior temporal lobe subregions. *The Journal of Neuroscience*, 36(5), 1490–1501. <https://doi.org/10.1523/JNEUROSCI.2999-15.2016>
- 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>
- Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2), 825–841. <https://doi.org/10.1006/nimg.2002.1132>
- Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, 62(2), 782–790. <https://doi.org/10.1016/j.neuroimage.2011.09.015>
- Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2002). Neural correlates of self-reflection. *Brain: A Journal of Neurology*, 125(8), 1808–1814. <https://doi.org/10.1093/brain/awf181>
- Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*, 14(5), 785–794. <https://doi.org/10.1162/08989290260138672>
- Kihlstrom, J. F., Albright, J. S., Klein, S. B., Cantor, N., Chew, B. R., & Niedenthal, P. M. (1988). Information processing and the study of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 145–178). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60226-9](https://doi.org/10.1016/S0065-2601(08)60226-9)
- Kihlstrom, J. F., Beer, J. S., & Klein, S. B. (2003). Self and identity as memory. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 68–90). Guilford Press.
- Kirby, D. M., & Gardner, R. C. (1972). Ethnic stereotypes: Norms on 208 words typically used in their assessment. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 26(2), 140–154. <https://doi.org/10.1037/h0082423>
- Kitayama, S., & Park, J. (2014). Error-related brain activity reveals self-centric motivation: Culture matters. *Journal of Experimental Psychology: General*, 143(1), 62–70. <https://doi.org/10.1037/a0031696>
- Koban, L., Gianaros, P. J., Kober, H., & Wager, T. D. (2021). The self in context: Brain systems linking mental and physical health. *Nature Reviews Neuroscience*, 22(5), 309–322. <https://doi.org/10.1038/s41583-021-00446-8>
- Koch, A., Alves, H., Krüger, T., & Unkelbach, C. (2016). A general valence asymmetry in similarity: Good is more alike than bad. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(8), 1171–1192. <https://doi.org/10.1037/xlm0000243>
- Kreisman, J. J., & Straus, H. (2021). *I hate you—Don't leave me: Understanding the borderline personality*. Penguin.
- Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10), 3863–3868. <https://doi.org/10.1073/pnas.0600244103>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(1), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Lambon Ralph, M. A. (2013). Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 369(1634), 1–11. <https://doi.org/10.1098/rstb.2012.0392>
- Leary, M. R., & Tangney, J. P. (2003). The self as an organizing construct in the behavioral and social sciences. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (2nd ed., pp. 1–18). The Guilford Press.
- Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of mechanical turk samples. *SAGE Open*, 6(1), 1–17. <https://doi.org/10.1177/2158244016636433>
- Linville, P. W. (1987). Self-complexity as a cognitive buffer against stress-related illness and depression. *Journal of Personality and Social Psychology*, 52(4), 663–676. <https://doi.org/10.1037/0022-3514.52.4.663>
- Ma, N., Baetens, K., Vandekerckhove, M., Kestemont, J., Fias, W., & Van Overwalle, F. (2014). Traits are represented in the medial prefrontal cortex: An fMRI adaptation study. *Social Cognitive and Affective Neuroscience*, 9(8), 1185–1192. <https://doi.org/10.1093/scan/nst098>
- Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, 11(1), Article 46. <https://doi.org/10.1038/s41467-019-13930-8>
- Maliniak, D., Powers, R., & Walter, B. F. (2013). The gender citation gap in international relations. *International Organization*, 67(4), 889–922. <https://doi.org/10.1017/S0020818313000209>
- Malle, B. F., & Horowitz, L. M. (1995). The puzzle of negative self-views: An exploration using the schema concept. *Journal of Personality and Social Psychology*, 68(3), 470–484. <https://doi.org/10.1037/0022-3514.68.3.470>
- Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, 35(2), 63–78. <https://doi.org/10.1037/0022-3514.35.2.63>

- Markus, H., & Kunda, Z. (1986). Stability and malleability of the self-concept. *Journal of Personality and Social Psychology*, 51(4), 858–866. <https://doi.org/10.1037/0022-3514.51.4.858>
- Markus, H., & Wurf, E. (1987). The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology*, 38(1), 299–337. <https://doi.org/10.1146/annurev.ps.38.020187.001503>
- Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2), 224–253. <https://doi.org/10.1037/0033-295X.98.2.224>
- Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist*, 20(3), 107–123. [https://doi.org/10.1207/s15326985ep2003\\_1](https://doi.org/10.1207/s15326985ep2003_1)
- Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, 97(5), 951–966. <https://doi.org/10.1037/a0028380>
- McConnell, A. R. (2011). The multiple self-aspects framework: Self-concept representation and its implications. *Personality and Social Psychology Review*, 15(1), 3–27. <https://doi.org/10.1177/1088868310371101>
- McConnell, A. R., Renaud, J. M., Dean, K. K., Green, S. P., Lamoreaux, M. J., Hall, C. E., & Rydell, R. J. (2005). Whose self is it anyway? Self-aspect control moderates the relation between self-complexity and well-being. *Journal of Experimental Social Psychology*, 41(1), 1–18. <https://doi.org/10.1016/j.jesp.2004.02.004>
- McConnell, A. R., & Strain, L. M. (2007). Content and structure of the self-concept. In C. Sedikides & S. J. Spencer (Eds.) *The self* (pp. 51–73). Psychology Press.
- Meiran, N., Diamond, G. M., Toder, D., & Nemets, B. (2011). Cognitive rigidity in unipolar depression and obsessive compulsive disorder: Examination of task switching, stroop, working memory updating and post-conflict adaptation. *Psychiatry Research*, 185(1–2), 149–156. <https://doi.org/10.1016/j.psychres.2010.04.044>
- Mitchell, J. P. (2009). Inferences about mental states. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 364(1521), 1309–1316. <https://doi.org/10.1098/rstb.2008.0318>
- Mitchell, S. M., Lange, S., & Brus, H. (2013). Gendered citation patterns in international relations journals. *International Studies Perspectives*, 14(4), 485–492. <https://doi.org/10.1111/insp.12026>
- Moran, J. M., Macrae, C. N., Heatherton, T. F., Wyland, C. L., & Kelley, W. M. (2006). Neuroanatomical evidence for distinct cognitive and affective components of self. *Journal of Cognitive Neuroscience*, 18(9), 1586–1594. <https://doi.org/10.1162/jocn.2006.18.9.1586>
- Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, 59(3), 2636–2643. <https://doi.org/10.1016/j.neuroimage.2011.08.076>
- North, R. J., & Swann, W. B., Jr. (2009). Self-verification 360: Illuminating the light and dark sides. *Self and Identity*, 8(2–3), 131–146. <https://doi.org/10.1080/15298860802501516>
- Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—A meta-analysis of imaging studies on the self. *NeuroImage*, 31(1), 440–457. <https://doi.org/10.1016/j.neuroimage.2005.12.002>
- Orth, U., & Robins, R. W. (2013). Understanding the link between low self-esteem and depression. *Current Directions in Psychological Science*, 22(6), 455–460. <https://doi.org/10.1177/0963721413492763>
- Orth, U., Robins, R. W., Trzesniewski, K. H., Maes, J., & Schmitt, M. (2009). Low self-esteem is a risk factor for depressive symptoms from young adulthood to old age. *Journal of Abnormal Psychology*, 118(3), 472–478. <https://doi.org/10.1037/a0015922>
- Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). Map making: Constructing, combining, and inferring on abstract cognitive maps. *Neuron*, 107(6), 1226–1238.e8. <https://doi.org/10.1016/j.neuron.2020.06.030>
- Price, A. R., Peelle, J. E., Bonner, M. F., Grossman, M., & Hamilton, R. H. (2016). Causal evidence for a mechanism of semantic integration in the angular gyrus as revealed by high-definition transcranial direct current stimulation. *The Journal of Neuroscience*, 36(13), 3829–3838. <https://doi.org/10.1523/JNEUROSCI.3120-15.2016>
- Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for research in the general population. *Applied Psychological Measurement*, 1(3), 385–401. <https://doi.org/10.1177/014662167700100306>
- Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1), 42–55. <https://doi.org/10.1038/nrn.2016.150>
- Rameson, L. T., Satpute, A. B., & Lieberman, M. D. (2010). The neural correlates of implicit and explicit self-relevant processing. *NeuroImage*, 50(2), 701–708. <https://doi.org/10.1016/j.neuroimage.2009.12.098>
- Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, 27(5), 709–748. [https://doi.org/10.1207/s15516709cog2705\\_2](https://doi.org/10.1207/s15516709cog2705_2)
- Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage*, 23(2), 752–763. <https://doi.org/10.1016/j.neuroimage.2004.06.035>
- Rosenberg, M. (1989). *Society and the adolescent self-image*. Wesleyan University Press.
- Schell, T. L., Klein, S. B., & Babey, S. H. (1996). Testing a hierarchical model of self-knowledge. *Psychological Science*, 7(3), 170–173. <https://doi.org/10.1111/j.1467-9280.1996.tb00351.x>
- Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, 91(6), 1402–1412. <https://doi.org/10.1016/j.neuron.2016.08.019>
- Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science*, 3(2), 102–116. <https://doi.org/10.1111/j.1745-6916.2008.00068.x>
- Seghier, M. L. (2013). The angular gyrus: Multiple functions and multiple subdivisions. *The Neuroscientist*, 19(1), 43–61. <https://doi.org/10.1177/1073858412440596>
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- 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>
- Shimamura, A. P. (2011). Episodic retrieval and the cortical binding of relational activity. *Cognitive, Affective & Behavioral Neuroscience*, 11(3), 277–291. <https://doi.org/10.3758/s13415-011-0031-4>
- Showers, C. (1992). Compartmentalization of positive and negative self-knowledge: Keeping bad apples out of the bunch. *Journal of Personality and Social Psychology*, 62(6), 1036–1049. <https://doi.org/10.1037/0022-3514.62.6.1036>
- Showers, C. J., Ditzfeld, C. P., & Zeigler-Hill, V. (2015). Self-concept structure and the quality of self-knowledge. *Journal of Personality*, 83(5), 535–551. <https://doi.org/10.1111/jopy.12130>
- Siegel, J. S., Power, J. D., Dubis, J. W., Vogel, A. C., Church, J. A., Schlaggar, B. L., & Petersen, S. E. (2014). Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Human Brain Mapping*, 35(5), 1981–1996. <https://doi.org/10.1002/hbm.22307>
- Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22(2), 189–228. [https://doi.org/10.1207/s15516709cog2202\\_2](https://doi.org/10.1207/s15516709cog2202_2)
- Sood, G., & Laohaprapanon, S. (2018). Predicting race and ethnicity from the sequence of characters in a name. ArXiv. <https://doi.org/10.48550/arXiv.1805.02109>
- Spalding, K. N., Jones, S. H., Duff, M. C., Tranel, D., & Warren, D. E. (2015). Investigating the neural correlates of schemas: Ventromedial prefrontal cortex is necessary for normal schematic influence on memory.



- The Journal of Neuroscience*, 35(47), 15746–15751. <https://doi.org/10.1523/JNEUROSCI.2767-15.2015>
- Stange, J. P., Alloy, L. B., & Fresco, D. M. (2017). Inflexibility as a vulnerability to depression: A systematic qualitative review. *Clinical Psychology*, 24(3), 245–276. <https://doi.org/10.1111/cpsp.12201>
- Steiger, A. E., Allemand, M., Robins, R. W., & Fend, H. A. (2014). Low and decreasing self-esteem during adolescence predict adult depression two decades later. *Journal of Personality and Social Psychology*, 106(2), 325–338. <https://doi.org/10.1037/a0035133>
- Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour*, 4(4), 361–371. <https://doi.org/10.1038/s41562-019-0800-6>
- Stopa, L., Brown, M. A., Luke, M. A., & Hirsch, C. R. (2010). Constructing a self: The role of self-structure and self-certainty in social anxiety. *Behaviour Research and Therapy*, 48(10), 955–965. <https://doi.org/10.1016/j.brat.2010.05.028>
- Summerfield, C., Luyckx, F., & Sheahan, H. (2020). Structure learning and the posterior parietal cortex. *Progress in Neurobiology*, 184, Article 101717. <https://doi.org/10.1016/j.pneurobio.2019.101717>
- Swann, W. B., Rentfrow, P. J., & Guinn, J. S. (2003). Self-verification: The search for coherence. In *Handbook of self and identity* (pp. 367–383). The Guilford Press.
- Swann, W. B., Jr., Wenzlaff, R. M., & Tatarodi, R. W. (1992). Depression and the search for negative evaluations: More evidence of the role of self-verification strivings. *Journal of Abnormal Psychology*, 101(2), 314–317. <https://doi.org/10.1037/0021-843X.101.2.314>
- Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, 22(3), 201–212. <https://doi.org/10.1016/j.tics.2017.12.005>
- Tarlow, E. M., & Haaga, D. A. F. (1996). Negative self-concept: Specificity to depressive symptoms and relation to positive and negative affectivity. *Journal of Research in Personality*, 30(1), 120–127. <https://doi.org/10.1006/jrpe.1996.0008>
- Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin*, 110(1), 67–85. <https://doi.org/10.1037/0033-2909.110.1.67>
- Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103(2), 193–210. <https://doi.org/10.1037/0033-2909.103.2.193>
- Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex*, 28(10), 3505–3520. <https://doi.org/10.1093/cercor/bhx216>
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136. [https://doi.org/10.1016/0010-0285\(80\)90005-5](https://doi.org/10.1016/0010-0285(80)90005-5)
- Unkelbach, C., Alves, H., & Koch, A. (2020). Negativity bias, positivity bias, and valence asymmetries: Explaining the differential processing of positive and negative information. In *Advances in experimental social psychology* (Vol. 62, pp. 115–187). Elsevier. <https://doi.org/10.1016/bs.aesp.2020.04.005>
- Vazire, S., & Wilson, T. D. (2012). *Handbook of self-knowledge*. Guilford Press.
- Wagner, D. D., Haxby, J. V., & Heatherton, T. F. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(4), 451–470. <https://doi.org/10.1002/wcs.1183>
- Wagner, I. C., van Buuren, M., Kroes, M. C., Gutteling, T. P., van der Linden, M., Morris, R. G., & Fernández, G. (2015). Schematic memory components converge within angular gyrus during retrieval. *eLife*, 4, Article e09668. <https://doi.org/10.7554/eLife.09668>
- Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods: Cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, 17(8), 513–523. <https://doi.org/10.1038/nrn.2016.56>
- Woolrich, M. (2008). Robust group analysis using outlier inference. *NeuroImage*, 41(2), 286–301. <https://doi.org/10.1016/j.neuroimage.2008.02.042>
- Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21(4), 1732–1747. <https://doi.org/10.1016/j.neuroimage.2003.12.023>
- Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, 14(6), 1370–1386. <https://doi.org/10.1006/nimg.2001.0931>
- Yazar, Y., Bergström, Z. M., & Simons, J. S. (2014). Continuous theta burst stimulation of angular gyrus reduces subjective recollection. *PLOS ONE*, 9(10), Article e110414. <https://doi.org/10.1371/journal.pone.0110414>
- Zeithamova, D., Mack, M. L., Braunlich, K., Davis, T., Seger, C. A., van Kesteren, M. T. R., & Wutz, A. (2019). Brain mechanisms of concept learning. *The Journal of Neuroscience*, 39(42), 8259–8266. <https://doi.org/10.1523/JNEUROSCI.1166-19.2019>
- Zhou, D., Cornblath, E. J., Stiso, J., Teich, E. G., Dworkin, J. D., Blevins, A. S., & Bassett, D. S. (2020). *Gender diversity statement and code notebook v1.0* [Computer software]. Zenodo.

Received June 17, 2021

Revision received April 11, 2022

Accepted April 14, 2022 ■