

Social, Neural, and Computational Mechanisms of Self-Evaluation and their Links to
Internalizing Symptoms

by

Moriah S. Stendel

A dissertation accepted and approved in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Psychology

Dissertation Committee:

Robert S. Chavez, Chair

Melynda D. Casement, Core Member

Jennifer H. Pfeifer, Core Member

John A. Clithero, Institutional Representative

University of Oregon

Spring 2025

© 2025 Moriah S. Stendel

DISSERTATION ABSTRACT

Moriah S. Stendel

Doctor of Philosophy in Psychology

Title: Social, Neural, and Computational Mechanisms of Self-Evaluation and their Links to Internalizing Symptoms

While most individuals possess a positively biased view of themselves, individuals who experience symptoms of internalizing disorders (e.g. anxiety, unipolar depression) often engage in *negative* self-evaluation. Negative self-evaluation is a core feature of internalizing disorders that predicts the severity and chronicity of symptoms. While a large literature has described such self-evaluative biases, less work has clarified the mechanisms that support them. As such, researchers have advocated for increased focus on integrating social, affective, and computational processes to provide a more mechanistic account of self-evaluation.

This dissertation draws on theory and methodology from social psychology, neuroscience, and computational psychiatry to identify possible mechanisms that link biased self-evaluation to internalizing disorders. Chapter I provides an introduction to self-evaluation and internalizing symptoms, emphasizing the importance of social factors, reward processes, and transdiagnostic considerations. Chapter II represents a review paper positioning *mentalizing* (i.e. inferring the beliefs of others) as a critical process driving negative self-evaluation. We describe different mentalizing strategies that might explain low self-esteem as a risk factor for internalizing disorders. Chapter III sought to build upon this work by exploring the neural mechanisms that contribute to self-esteem within rich social contexts. Here, we looked at the similarity in neural representation when an individual reflects on themselves and when their

peers reflect on them. We replicate findings that show that the self is recapitulated in the brains of friends (Chavez & Wagner, 2020), but that divergent representations in the MPFC, which is associated with both self- and evaluative processing, predict higher self-esteem. Chapter IV aims to better understand the cognitive mechanisms underlying self-evaluations in a social context using drift-diffusion modeling. This chapter demonstrates that the relationship between internalizing symptoms and altered self-evaluation is best explained by changes in the accumulation and integration of evidence in support of positive self-views, and to a lesser degree, negative self-views. Chapter V integrates these findings and links them to broader theory and research on self-enhancement and clinical treatment while exploring limitations and future directions. Together, the work presented in this dissertation shows that thinking about ourselves in social contexts, either imaginally (Chapter II), alongside real-life peers (Chapter III), or in direct social comparison (Chapter IV), can shift our sense of self, in ways that can be beneficial or detrimental to mental health.

This dissertation includes unpublished co-authored material.

ACKNOWLEDGEMENTS

The work contained in this dissertation reflects a collective effort. It represents a series of unexpected connections, paths I never thought would lead anywhere, yet here I am, thanks to the benevolent support of many.

I am forever grateful to my advisor, Rob Chavez, for his advocacy, generous support, and formative influence on my thinking. Whenever I presented Rob with the tangled threads of an idea, he took the time to unwind them through our round-about and ever-interesting conversations. He critically contributed to the technical part of these projects but also helped me situate them in the bigger picture. During my bouts of imposter syndrome, Rob always helped tune my ‘sociometer’ – or, would caringly remind me that we are *all* imposters anyway, freeing me up to enjoy the ride. Thank you for celebrating my unconventional skillset and patiently supporting me through the challenges of working in interdisciplinary spaces.

I am also beyond lucky and humbled for the support of my dissertation committee, Melynda Casement, John Clithero, and Jennifer Pfeifer, experts in this field who so generously shared their knowledge and feedback. Thank you to Jenn for her wisdom on the relationship between social cognition and the self, Melynda’s remarkable wealth of knowledge in clinical science, and John’s insights into drift-diffusion modeling as well as his saint-like patience in helping me understand and execute it. Thank you to my clinical supervisors over the years for passing on wisdom that shaped this dissertation, with special mention to Melynda (once again), Ruth Ellingsen, Jenny Mauro, Karrie Walters, and Maureen Zalewski. Without Lori Olson, I would have literally been unable to get through this program, and am deeply grateful for the practical support she has provided me and so many others. And thank you to Mathieu Landry,

for his risk in taking me on as a research assistant at McGill and his subsequent mentorship; Mathieu's belief in me propelled me on this path that I had originally ruled out as impossible.

Thank you to the research participants who have offered time, energy, and vulnerability in support of this work. None of this would have been possible without our lab's current and former graduate students, postdoctoral scholars, research staff, and affiliates, including Sam Chavez, Jack Kapustka, Taylor Guthrie, Rachel Jacobson, Sonny Mattek, and Dale Tovar, as well as a thoughtful, brilliant team of research assistants and staff: Maddie Alms, Youri Benadjaoud, Faith Collins, Ryan Conaghan, Austin Frost, Hannah Grant, Conor Hollowell, Louis Kerner, Clara Morrison, Debbie Okeke, Maddie Schall, and Mads Young. Thank you all *so much*, I am truly honored to have been a node in this web of brilliant and kind humans.

Thank you to my fellow graduate students, friends, co-op mates, loved ones near & far, and the Oregon wilderness, all of whom have supported me in critical ways and allowed me to build a rich and beautiful life here that supported this work.

Finally, I am beyond grateful to my family for providing me with such a strong foundation of love and support. They have unquestionably supported all my dreams, including the 'PhD' one that took me far away from home in pursuit of knowledge and adventure. I promise to pay forward all this generosity as best as I can.

TABLE OF CONTENTS

I. GENERAL INTRODUCTION	12
Self-Evaluation.....	12
Self-Evaluation is Social and Affective	15
Social Dimensions of Self-Evaluation.....	16
Affective Dimensions of Self-Evaluation.....	18
Transdiagnostic Considerations	21
Overview of Subsequent Chapters	22
II. THE ROLE OF META-PERCEPTION IN SELF-ESTEEM AND IMPLICATIONS FOR MENTAL HEALTH: A REVIEW	24
Introduction	24
Interpersonal Theories of Self-Esteem.....	25
Meta-Perception as a Theoretical Commonality across Interpersonal Theories of Self-Esteem	28
Individual Differences in Meta-Perception Processes	30
Propensity	31
Accuracy	34
Permeability.....	38
Internalizing Symptoms as a “Perfect Storm” of Vulnerability.....	43
Summary of Mentalizing Research across Internalizing Symptoms	45
Meta-Perception in Internalizing Disorders	47
Therapeutic Implications and Treatment Targets.....	51
Conclusions and Future Directions	55
III. SELF-ESTEEM MODULATES THE BRAIN-TO-BRAIN SIMILARITY OF THE REPRESENTATION OF THE SELF IN THE BRAINS OF OTHERS	58

Introduction	58
Methods	62
Participants	62
Procedures	63
Behavioral Measures.....	63
Neuroimaging Procedures.....	64
Image Acquisition.....	66
Transparency and Openness	66
Analyses	66
Functional Localizer Preprocessing and Analysis.....	66
Round-Robin Task Preprocessing and Response Pattern Estimation	67
Multivariate Round-Robin Task and Self-Esteem Analyses.....	68
Results	70
Replication of the Neural Self-Recapitulation Effect.....	70
Relationship of Self-Recapitulation Strength and Self-esteem	71
Discussion	73
Limitations and Conclusions	75
IV. COMPUTATIONAL MECHANISMS OF SELF-ENHANCEMENT DURING SOCIAL COMPARISON AND THEIR RELATIONSHIP TO INTERNALIZING SYMPTOMS	78
Introduction	78
Methods	83
Sample	83
Measures	84
Procedures	87
Analysis	88

Data Analytic Plan	89
Results	92
Descriptive Statistics	92
Mean Response Times	93
Endorsement Rates	96
Mixed Effects Models	97
Hierarchical Bayesian Drift-Diffusion Modeling.....	99
Discussion	103
V. GENERAL DISCUSSION	110
Theoretical Implications.....	112
Self-Positivity	112
Treatment Considerations.....	114
Beyond the Self	116
A Middle Path.....	117
Limitations and Future Directions.....	119
Conclusion.....	125
REFERENCES CITED.....	126

LIST OF FIGURES

Figure	Page
2.1 Meta-Perception Feedback Loop in Internalizing Disorders	51
3.1 Results from the Multi-Group Replication of the Self-Recapitulation Effect	71
3.2 Correlation of Self-Esteem and Recapitulation Effect within the MPFC.....	72
4.1 Schematic of the Trait-Evaluation Task	85
4.2 Schematic of the Drift Diffusion Model	92
4.3. Average Response Times and Endorsement Across Participants.....	94
4.4 Mean Response Times Plotted Across Word Valence	100
4.5 Drift Rate by Individual Difference Measures.....	103
4.6 Correlation Matrix of Internalizing Symptoms and DDM Parameters.....	105

LIST OF TABLES

Table	Page
2.1 Meta-perceptive Accuracy, Permeability, and Propensity: Description and Examples	42
4.1 Mean RTs and Endorsement Rates by Valence and Response.....	94
4.2 Linear Mixed Effects Models Predicting RTs	101
4.3 Drift-Diffusion Model Fits.....	102

GENERAL INTRODUCTION

Each individual possesses inherent value and dignity, however many fail to experience themselves as worthy, increasing risk for a variety of psychological disorders. The present dissertation will focus on factors that influence how individuals evaluate themselves, and how variations in this process relate to psychiatric symptoms. In particular, the presented research will make a case for studying how social and reward processes influence self-evaluation and can give rise to transdiagnostic risk factors for internalizing disorders.

Self-Evaluation

Self-evaluation stands as a fundamental aspect of human cognition, with evidence of its emergence early in the archeological record through artifacts such as jewelry (Leary et al., 2007) or handprint tracings in the Pettakere caves as prehistoric forms of self-portraits. While our ability to inhabit the phenomenology of ancient humans or other species is limited, there is value in thinking about the origins of self-evaluation – a behavior that likely evolved for the many benefits it can confer, such as planning for our futures, reflecting on our pasts, changing our behaviors, and acting with intention. This same process, however, is also the source of so much human suffering, some of which this dissertation hopes to better understand.

Around 44,000 years after the Pettakere Cave handprints, William James described the self as both the *I*, the subjective, first-person perceiver, and the *me*, the third-person object of perception, a theory I believe is illustrated in those ancient handprints (James, 1890). In depicting both the person who left the handprint (the *I*) and the representation of themselves they sought to portray (the *me*), these prints capture the reflexive nature of self-evaluation, wherein

one assumes both the roles of judge and the judged, knower and the known (Kaplan, 1986). Here, self-evaluation is like any other evaluative process, where meaning is attached to an object, however, the subject doing the evaluating happens to also be the very thing being evaluated (Tesser, 2003). This reflexivity underscores the significance of *self-feelings* (Kaplan, 1986) – emotions felt about oneself – which both precede and result from self-evaluation.

The combination of ‘self-evaluation’ and ‘self-feelings’ can be understood as *self-esteem*. Self-esteem is an individual’s subjective evaluation of their own worth and encompasses both cognitive and affective dimensions, including beliefs about the self (e.g. “I am good”) and affective states (e.g. shame, pride; Tangey, 2002), either globally (i.e. trait self-esteem, Rosenberg et al., 1995) or in response to transient factors (i.e. state self-esteem; Heatherton & Policy, 1991). In other words, self-esteem is “affectively laden self-evaluation” (Leary & MacDonald, 2003). Self-esteem and self-evaluations are so tightly coupled that some work has questioned their discriminant validity (e.g. Porter et al., 2016). While more work is necessary to tease apart these constructs, at the least, we can acknowledge that self-evaluations are critical to how we understand and experience our own worth.

Alterations in self-evaluation are also a core feature of internalizing disorders, a descriptive label capturing diagnoses that feature persistent negative emotions directed inward, rather than outwardly expressed, such as in unipolar mood and anxiety disorders (Eaton et al., 2013). Negative self-evaluation is listed as a symptom of internalizing disorders across multiple diagnostic classification systems. For example, the ICD-10 and DSM-5 include disruptions to self-referential processing as a symptom of major depression (e.g. “feelings of worthlessness”, “loss of confidence or self-esteem”, “unreasonable feelings of self-reproach”), persistent depressive disorder (e.g. “low self-esteem”), premenstrual dysphoric disorder (e.g. “self-

deprecating thoughts”), and social anxiety disorder (e.g. “fear of negative evaluation”) (DSM 5, APA, 2013; ICD, WHO, 2021).

Beyond diagnostic categorization, the link between negative self-evaluation and clinical symptoms has been central to prevailing theoretical models of internalizing disorders. For example, Beck’s (1967) seminal cognitive model of depression emphasized the significant role of negative self-evaluations. In this model, automatic, negative thoughts about oneself, the world, and future interact to contribute to the development and maintenance of depression. This model remains essential to contemporary clinical research and practice, extending beyond depression to link negative self-evaluation with a variety of psychological disorders (see: Beck & Haigh’s *Generic Cognitive Model*, 2014; Clark & Well’s *Cognitive Model of Social Phobia*, 2005; etc.).

Empirical work also highlights the increased self-focus and negative self-beliefs characteristic of internalizing disorders (Ginot, 2012; McEvoy et al., 2019; Mor & Winquist, 2002). Various studies demonstrate elevated self-referential processing in depression (Kuiper & Derry, 1982; Lemogne et al., 2012; Li et al., 2017; Nejad et al., 2013; Northoff & Tumati, 2019), social anxiety disorder (Abraham et al., 2013; Lin et al., 2018; Yoon et al., 2019), and generalized anxiety disorder (Cui et al., 2022), particularly *negative* self-referential processing such as rumination and self-criticism (Dinulescu et al., 2021; Ke et al., 2020; Kircanski et al., 2015; Nolen-Hoeksema et al., 2008). While internalizing disorders represent heterogeneous conditions with diverse symptom profiles within disorders, negative self-evaluation appears to exert a significant influence on severity and prognosis. For example, negative self-evaluation represents the cognitive bias most strongly predictive of depression severity (Orchard & Reynolds, 2018), increases the length and frequency of depressive episodes (Nolen-Hoeksema et

al., 2008), alone can categorize whether an adolescent has clinical depression (Orchard & Reynolds, 2018), and predicts the severity of social anxiety (Moscovitch et al., 2009). Negative self-evaluation also mediates the relationship between internalizing and externalizing disorders (Lee & Stone, 2011). Neuroimaging research has demonstrated aberrant activity in brain regions and networks involved in self-processing in internalizing disorders (Lemogne et al., 2009; Lin et al., 2018; Nejad et al., 2013; Northoff & Tumati, 2019), lending further support that elevated, negative self-focus is one factor that contributes to internalizing symptomatology. Further work must continue to study self-evaluation as a critical, transdiagnostic process in clinical disorders.

Self-Evaluation is Social and Affective

While self-, social-, and reward-processing are often studied separately, emerging work has attempted to integrate research across these areas (e.g. Hobbs et al., 2021; Luyten & Fonagy, 2018; Pfeifer & Allen, 2021).

Independently, widespread social disruptions are observed in internalizing disorders. Social withdrawal, either due to fear or anhedonia, is common across internalizing disorders (Rappaport & Barch, 2020), along with reductions in social support (Rueger et al., 2010), increased rejection sensitivity (Heeren et al., 2017; Harrewijn et al., 2017; Hsu et al., 2015; Jankowski et al., 2018; Kumar et al., 2017; Silk et al., 2014), and socio-cognitive deficits (Dainer-Best et al., 2017; Gotlib & Joormann, 2010). Internalizing disorders also feature broad disruptions in reward processing (Kujawa & Burkhouse, 2017), such as hypersensitivity to punishment/negative stimuli (Eshel & Roiser, 2010; Ng et al., 2019), hyposensitivity to reward/positive stimuli (Bylsma et al., 2008; Eshel & Roiser, 2010; Feeser et al., 2013; Forbes & Dahl, 2012; Luking et al., 2016; Moran et al., 2012), and reduced motivation to pursue rewarding

stimuli (Bryant et al., 2017), with these disruptions predicting the chronicity of symptoms (Vrieze et al., 2013).

Increasingly, research is highlighting how self-, social- and reward- processes interact to maintain internalizing symptoms, along with the importance of examining this interaction. For instance, Hobbs et al. (2021) show that while depression symptoms cannot be predicted by independent alterations in self-referential, emotion, or reward processing, they are strongly predicted by performance on tasks that require the coordination of these three domains. Other work has shown that depressive symptoms are associated with attenuated neural sensitivity to positive social feedback, above and beyond monetary rewards, which are more commonly used in reward research (Ait Oumeziane et al., 2019). Internalizing symptoms are linked to attenuated reward response to self-relevant stimuli (Northoff & Hayes, 2011; Quevedo et al., 2017) and positive social stimuli (Davey et al., 2011; He et al., 2019; Zhang et al., 2020; Zhang et al., 2017), with preferential attention towards negative and reduced attention towards positive social stimuli (Disner et al., 2017; Rantanen et al., 2021). As such, there have been calls for increased attention to processes that lie at the intersection of self, social, and reward processes (Luyten & Fonagy, 2018), however, previous work often fails to provide a granular understanding of social and reward- processes that are entwined with self-evaluation.

Social Dimensions of Self-Evaluation

Self-evaluation often transpires in rich social contexts. This notion is not new – in 1909 Cooley wrote, “Social consciousness, or awareness of society, is inseparable from self-consciousness because we can hardly think of ourselves except with reference to a social group of some sort, or of the group except with reference to ourselves . . . Self and society are twin-

born, we know one as immediately as we know the other, and the notion of a separate and independent ego is an illusion.” Indeed, the self is not static and does not form in a vacuum; it is constructed over time in relation to contextual and interpersonal encounters (Lieberman & Pfeifer, 2004). Early theories on the development of self underscore the reciprocal relationship between social factors and self-concept, via social interactions, feedback from others, and the internalization of societal norms and values (e.g. *Symbolic Interactionism*; Mead, 1934). Self-evaluations are measured against a backdrop of norms and expectations, occur alongside peer evaluations (Swallow & Kuiper, 1988), emerge in the context of social comparisons (Festinger, 1954), integrate feedback from others, serve social motives (Sedikides & Skowronski, 2009), and are related to sociocultural identity (Pfeifer & Peake, 2012). When considering internalizing disorders, social factors might be especially relevant given the role of social comparison (McCarthy & Morina, 2020) and altered processing of social feedback (Caouette & Guyer, 2016; Shany et al., 2022) in the development and maintenance of clinical symptoms.

Moreover, self- and social- evaluations recruit shared cognitive and neural substrates, suggesting that these are likely overlapping processes. Individuals evaluate themselves using the inferred beliefs of others (Jankowski et al., 2014; Pfeifer & Peake, 2012; Pfeifer et al., 2009) and evaluate others using egocentric frames of reference (Todd & Tamir, 2024; Wagner et al., 2012). Self- and other- are also tangled in the brain, where these representations share common real estate across social-affective brain networks, especially the medial prefrontal cortex (MPFC; Chavez, 2020; Wagner et al., 2012). While a debate still continues about whether the self is a special knowledge structure or just another form of person-perception, the relationship between self and peer perception motivates further research that considers social dimensions of self-evaluation.

Despite clear evidence for social dimensions of self-evaluation it is typically depicted as a private, interior process. For example, an overwhelming amount of research on self-evaluation uses the Self-Referent Encoding Task (SRET; Derry & Kuiper, 1981), which presents participants with a trait-adjective and asks them to respond YES or NO as to whether the trait describes them. While this seminal task has afforded much of our understanding of self-referential processing biases in internalizing disorders, its design ignores that self-evaluation frequently occurs in social contexts. The present dissertation aims to situate self-evaluation as a socio-cognitive process by studying self-referential processing within the context of social comparisons or real-life social relationships, and by connecting disparate findings from social and clinical psychology.

Affective Dimensions of Self-Evaluation

It is also critical to study self-evaluation within a broader framework of affective processes, such as reward, motivation, and valence, since what distinguishes self-evaluation from self-knowledge is its affective dimension. Self-evaluations consist of more than just knowledge about ourselves but also our attitudes and emotions about this knowledge (Beer, 2012). Since this affective component is critical to self-evaluation, we must consider how the self connects to reward and valence.

Firstly, we can consider self-evaluation as a motivated behavior that moves us towards desired states and away from undesired states (Kaplan, 1986). Researchers have highlighted different motives for self-evaluation, from self-assessment (to know ourselves), self-improvement (to make ourselves better), self-verification (to be seen as we are), to self-enhancement (to feel positive about ourselves; Gregg et al., 2011; Sedikides & Strube, 1995; Wayment & Taylor, 1995).

We can also consider how reward processes inform self-evaluation, which are often understood as intertwined processes (Yankouskaya et al., 2020). Self-evaluation recruits reward systems in the brain (Chavez et al., 2022; Chavez, 2017; de Greck et al., 2008; Northoff & Hayes, 2011) and shows overlapping neural substrates with positive stimuli (Chavez et al., 2017). These very same motivational and reward systems are broadly disrupted in internalizing disorders (Forbes, 2009; O'Callaghan & Stringaris, 2019; Weinberg, Liu, Hajcak & Shankman, 2015; Nelson et al., 2016) and these disruptions likely exert influence on how self-related stimuli are processed in individuals experiencing internalizing symptoms.

Research on self-evaluation in internalizing disorders has predominantly focused on negativity, with work on positive valence systems lacking (Vanderlind et al., 2020). While the structure of valence remains an ongoing debate (see Mattek et al., 2017), there is some consensus that positive and negative valence systems represent two, independent dimensions rather than a ‘bipolar’ construct (Briesemeister et al. 2020). This suggests that both negativity and lack of positivity might uniquely contribute to symptoms. Emerging work is highlighting the role of reduced self-positivity, independent of negativity, in internalizing symptoms (Collins et al., 2023; Goldstein et al., 2015). For instance, internalizing disorders are associated with attenuated reward responses in response to self-relevant stimuli (Quevedo et al., 2017) and positive social feedback (Caouette & Guyer, 2016; Shany et al., 2022). Depressed individuals are also more likely to endorse fewer positive than negative self-evaluations (Collins et al., 2023) and lack positive memory biases observed in non-depressed groups (Burt et al., 1995; Dillon, 2015a). Together, this growing body of work suggests that a “cognitive immunization” towards positive, self-relevant information contributes to symptoms above and beyond negative self-evaluation (Kube 2019). When we consider that blunted positivity is broadly implicated in internalizing

disorders (e.g. Bress et al., 2012; Bylsma et al., 2008; Forbes et al., 2010; Proudfit, 2015; Reichenberger et al., 2017; Vanderlind et al., 2020; Watson, Clark, & Carey, 1988), it logically follows that this anhedonic effect would extend to self-relevant stimuli, however further work is needed in this area.

Luckily, we can draw on the extensive body of non-clinical literature linking positivity biases to well-being. While one might intuit that accurate perception supports welfare, there is considerable evidence that having ‘positive illusions’ (Taylor & Brown, 1988) about ourselves and the world around us improves our mental health (Alessandri et al., 2014; Dufner et al., 2018; Schütz & Baumeister, 2017). Most relevant here is the *self-enhancement effect* (Krueger, 1998), the pervasive tendency for humans to overestimate their own abilities and positive qualities. This effect is one of the most consistent and robust in psychological research, where we find that humans have a general tendency to inflate their self-views across a variety of domains (Alicke & Sedikides, 2011; Hohman & Brown, 2020; Krueger, 1998; Leary, 2007; Zell et al. 2020). The extent and variety of self-enhancing strategies are so numerous and varied that they have been deemed the “self-zoo” (Tesser, Martin, & Cornell, 1996).

Despite robust evidence supporting the self-enhancement bias, it is crucial to note that not everyone self-enhances. In some cases, this can be adaptive. Self-criticism can serve the function of self-protection or eliciting support (Leary, 2007). Humility supports healthier social relationships (Tongeren, Davis, & Hook, 2014) and pro-social behaviors (Exline & Hill, 2012; LaBouff, Rowatt, Johnson, Tsang, & Willerton, 2011). Grappling with our flaws or less ideal traits can set the stage for healthy self-reflection and improved interpersonal outcomes. However, at the extreme, negative self-views can also reach a level of clinical severity, where the lack of positive self-evaluation ceases to be adaptive.

The absence of self-enhancement observed in internalizing disorders does not always necessitate an opposing, negative bias. For instance, research on depressive realism suggests that the negative cognitions observed in depression may reflect a more accurate assessment of the world, albeit one that lacks the positive bias crucial for supporting psychological well-being (Alloy & Abramson, 1988; Moore & Fresco, 2012). One recent study failed to replicate earlier findings supporting this theory (Dev et al., 2022), motivating the need for continued work in this area that considers whether and how possible biases in the processing of both positive and negative stimuli can contribute to symptomatology.

The present dissertation aims to enhance our understanding of affect in self-evaluation, considering both positive and negative biases and their relationship to well-being. To this end, each chapter bridges work from social psychology, which emphasizes the intrapersonal benefits of self-enhancement, and clinical psychology, which focuses on the mental health costs that emerge from negative self-evaluation. Moreover, each paper acknowledges the role of valence and reward systems in self-evaluation, either by investigating neural systems involved in both self- and evaluative processing, examining how the presence of negative self-evaluation and absence of positive self-evaluation both contribute to internalizing symptoms, or theoretically integrating literature about reward- and self-processing.

Transdiagnostic Considerations

This dissertation attempts to clarify the role of self-esteem as a transdiagnostic process that is typically studied in social psychology but can provide insights into clinical phenomena. Though diagnostic categories have played a longstanding role in our understanding of psychopathology, there is emerging consensus around the need to consider transdiagnostic

processes common to multiple disorders to better integrate our understanding of mental health disorders, explain high rates of comorbidity, and improve the efficacy of treatments (Low et al., 2021).

Low self-esteem is a transdiagnostic factor common to multiple psychiatric disorders, including depression (e.g. Kernis et al., 1998; Orth & Robins, 2013), anxiety (e.g. Koban et al., 2017; Sowislo & Orth, 2013), substance use disorder (e.g. Silverstone & Salsali, 2003), borderline personality disorder (e.g. Winter et al., 2018), body dysmorphic disorder (e.g. Kuck et al., 2021), and eating disorders (e.g. Colmsee et al., 2021). Not only is low self-esteem commonly experienced across many mental health diagnoses, but low self-esteem also connotes a strong risk for the development of later mental illness (Henriksen et al., 2017; Orth et al., 2009; Trzesniewski et al., 2006). For instance, a large meta-analysis shows a robust, directional link between self-esteem and depression, supporting that low self-esteem creates vulnerability for the development of later depression rather than alternate models, like the “scar” model (depression erodes self-esteem) or diathesis-stress models (Orth & Robins, 2013). Given that self-esteem represents a key transdiagnostic process, recent papers have highlighted the need to better understand self-esteem to advance our understanding of psychiatric conditions (Hsu et al., 2021; Low et al., 2021). This dissertation intends to heed this call by considering the transdiagnostic impact of self-esteem across internalizing disorders.

Overview of Subsequent Chapters

The research conducted across the three chapters of this dissertation was designed to provide a focused inquiry into how disruptions in self-evaluation, occurring in social contexts, can give rise to transdiagnostic risk factors for internalizing symptoms. While it is well-

established that self-evaluation is linked to internalizing symptoms, the present dissertation seeks to examine this relationship from multiple angles to increase our understanding of their interplay. The following work hopes to clarify the role of both positive and negative self-evaluation as they relate to internalizing symptomatology and related transdiagnostic constructs, like self-esteem. Moreover, the research presented in this dissertation situates self-evaluation as a sociocognitive process, by studying self-evaluation in the context of real social groups and social comparisons. Finally, the present work aims to leverage methodological advances in computational modeling and neuroimaging to provide a more granular understanding of how cognitive and neural processes give rise to aberrant self-evaluation.

Chapter I of this thesis is a theoretical review intended to update and expand our understanding of socio-cognitive processes that link self-evaluation and self-esteem and to further identify progress and gaps in this field of research. Following this review is a set of two interrelated and novel empirical analyses (Chapters III and IV) using computational methods to understand self-evaluation within social contexts. This dissertation contains previously published or under review and co-authored material. Chapter II is currently in revision and was co-authored with R. Chavez. Chapter III is currently in revision and was co-authored with T. Guthrie, V. Guazzelli Williamson, and R. Chavez. Chapter IV was co-authored with R. Chavez and J. Clithero.

CHAPTER II

THE ROLE OF META-PERCEPTION IN SELF-ESTEEM AND IMPLICATIONS FOR MENTAL HEALTH: A REVIEW

This work is in revision and was co-authored with R. Chavez. I was the lead author of this publication and responsible for the writing of the manuscript with input and assistance from the co-author.

Introduction

Low self-esteem is a critical transdiagnostic factor that plays a prominent role in the onset and maintenance of various psychiatric disorders, such as depression (e.g., Kernis et al., 1998; Orth & Robins, 2013), anxiety (e.g. Koban et al., 2017; Sowislo & Orth, 2013), substance use disorder (e.g., Silverstone & Salsali, 2003); borderline personality disorder (e.g., Winter et al., 2018), body dysmorphic disorder (e.g., Kuck et al., 2021) and eating disorders (e.g., Colmsee et al., 2021). Critically, low self-esteem is one of the most significant risk factors for suicidality (Brausch & Decker, 2014; Overholser et al., 1995; Wild et al., 2004).

Though diagnostic categories have played a longstanding role in our understanding of psychopathology, there is emerging consensus around the need to consider transdiagnostic processes common to multiple disorders in order to better integrate our understanding of mental health disorders, explain high rates of comorbidity, and improve the efficacy of treatments (Low et al., 2021). Given that self-esteem represents a key transdiagnostic process, recent papers have

highlighted the need to better understand self-esteem to advance our understanding of psychiatric disorders (T.-Y. Hsu et al., 2021; Low et al., 2021; Perinelli et al., 2022).

Historical theories of self-esteem have depicted it as fundamentally interpersonal, whereby our view of ourselves relates to our social relationships. Despite positing social factors as a critical aspect of self-esteem, few theories clarified the mechanisms that allow for social factors to influence our self-views. Here, we propose that *meta-perceptions* – the beliefs we hold about how others perceive us – are one cognitive process that allows for the views of others to inform our self-esteem. We will provide a non-exhaustive view of dominant theories in the self-esteem literature with the aim of illuminating the role that meta-perceptions play in self-esteem. Critically, we propose that individual tendencies in forming meta-perceptions can differentially impact self-esteem.

Interpersonal Theories of Self-Esteem

Common definitions of self-esteem typically depict it to be a private, interior process, defined as a personal evaluation of one's own worth, or the degree of intrinsic satisfaction one has with one's self. Indeed, we do have privileged access to our own inner world, and introspecting this private knowledge might be the best strategy to know and evaluate ourselves (Sedikides & Skowronski, 1995). However, such intrapersonal definitions often fail to capture our own phenomenological experience of self-esteem, which does not always feel like a solitary, inner process but instead can seem highly responsive to our social world, perhaps flowing after receiving positive feedback or ebbing with romantic rejection. A century of theorizing has circled around this intuition that our self-esteem is tightly woven with our social relationships,

producing many variations on the main idea that self-esteem is not merely a personal evaluation of our worth but is at least in part a reflection of how we believe others perceive and evaluate us.

This connection was theorized as early as Cooley's *looking-glass self* (1902), which states that we see ourselves through the eyes of others and use this information to construct self-representations. He described the Self to consist of "three principal elements: the imagination of our appearance to the other person, the imagination of his judgment of that appearance, and some sort of self-feeling, such as pride or mortification" (Cooley, 1902, p. 152). Cooley was considered an early proponent of symbolic interactionism, a framework that suggests that meaning emerges from social interaction with others and society, and thus one's self will be constructed through such interactions. Later iterations of symbolic interactionism took a more deterministic stance, positing that the Self can *never* be experienced directly but only indirectly via the perspective of other individuals or the generalized perspective of collectives (Mead, 1934).

Later in the century, Kinch (1964) formalized this view into a causal model containing four variables: The individual's self-concept (S), her perception of the responses of others (P), the actual responses of others (A), and her behavior (B). He posited that $A \rightarrow P \rightarrow S \rightarrow B$, or in other words, other people's appraisals influence our perception of what we think others think of us, which goes on to influence our self-representations and subsequent behaviors. Shrauger & Schoeneman (1979) codified the terminology of direct, reflected, and other appraisals, which map onto Kinch's S , P , and A variables accordingly. *Sociometer theory* (Leary & Baumeister, 2000) roots Kinch's causal model in an evolutionary framework by suggesting that self-esteem evolved to monitor our relational value so that we can adjust future behavior to confer the evolutionary benefits of group inclusion. On this view, our self-esteem fluctuates in accordance

with our social approval with our feelings about ourselves serving as a gauge of social acceptance and rejection.

We can also consider self-esteem through *self-enhancement theory* (Krueger, 1998; this bias goes by many names including *better-than-average effect*, *overconfidence*, *self-inflation*, and many more), which suggests that self-esteem is maintained by an inherent motivation to view oneself in a positive light, even at risk of ignoring or minimizing negative peer-perceptions. This effect is one of the most consistent across psychological research where we find that humans have a general tendency to inflate their self-views across a variety of domains (Alicke & Sedikides, 2011; Hohman & Brown, 2020; Krueger, 1998). In contrast to theories of self-enhancement, Swann's self-verification theory (Swann & Read, 1981) suggests that self-esteem is not dependent on being perceived positively by our peers, but rather by being perceived accurately.

At first glance, these theories appear competing and are often positioned as such – for instance, while some suggest that accurate self-perception connotes higher self-esteem, others pose that positive (even inaccurately so) self-perception supports higher self-esteem. However, unified across these models is the notion that we selectively incorporate social information into our self-concept to inform our self-esteem. By considering the possibility that individuals use (or ignore) social appraisals to maintain self-esteem, we can bridge clashing theories – for instance, perhaps positive self-esteem is bolstered by a flexible strategy of incorporating flattering information from others (symbolic interactionism) while ignoring other cues that might paint us in a negative light (self-enhancement). Across a whirlwind of theories, we see variations on a theme: self-esteem emerges in conversation with the social environment and requires the tracking, evaluating, and selective filtering of social information.

Notably, interpersonal theories of self-esteem do not always clearly account for mechanisms that allow for social factors to influence our self-views. How do the evaluations others hold of us get from their minds to ours to shape our sense of self? Although we occasionally receive direct feedback from others, more often than not we are inferring their mental states to make our best predictions about how they might perceive us. Interpersonal theories of self-esteem imply the existence of some cognitive process that allows for the interpretations of others to get mixed in the fold of our self-views, yet they rarely make explicit the unstated assumption that meta-perceptive processes might partially underlie self-esteem.

Meta-Perception as a Theoretical Commonality across Interpersonal Theories of Self-Esteem

Kenny (1994) writes, “We perceive [others], they perceive us in return. Not surprisingly, then, people try to perceive others’ perceptions of them,” describing the very human pull to form *meta-perceptions*¹, or beliefs about how others perceive us. Meta-perception falls under the umbrella of *mentalizing* – the process of inferring mental states such as emotions, thoughts, or intentions that underlie our behaviors or the behavior of others (Frith & Frith, 2006). Our capacity to mentalize allows us to meet the demands of ever-changing, complex social environments that require the flexible coordination of behavior based on shared understanding (Tomasello & Vaish, 2013). Meta-perception is a form of mentalizing process that allows us to create inferences about what others think of us, requiring individuals to draw from a broad array of socio-cognitive processes such as perspective-taking, empathy, and theory of mind. Meta-

¹ Another term frequently used to describe meta-perceptions is “reflected appraisals.” For parsimony, this paper will use the former term, however it should be noted that these terms are used interchangeably in the literature.

perceptions consist of a *perceiver*, a *target*, and a specific type of *content*, for instance, “I (the perceiver) think that my cousin (the target) thinks I am kind (content)”, or “I (the perceiver) believe that my classmates (the target) think I am mean (content)”, where the target can be a specific individual or a generalized collective. Meta-perceptions can be formed moment-to-moment or learned from accumulated data over longer time periods; by information communicated explicitly (e.g. direct feedback) or implicitly (e.g. body language, facial expressions, tone of voice, contextual implications, or behaviors); or can even be data-free speculations (Luyten & Fonagy, 2015).

While interpersonal theories of self-esteem lean heavily on the notion that the views of others inform our self-views, they do not always clearly account for the process that bridges these two sets of perceptions. Some have proposed meta-perceptions as a candidate mechanism to describe the process in which others’ appraisals inform our self-view (Pfeifer & Peake, 2012). Rather than using introspection or direct feedback to form self-impressions, individuals use the impressions of others to shape how they see themselves. This requires the recruitment of mentalizing processes to interpret incomplete and ambiguous social cues to form third-person, inferred impressions of what others think of us — i.e. *meta-perceptions*. Across varying time scales these meta-perceptions come to inform how we come to see ourselves, harking back to Kinch’s (1963) causal model of self-concept.

Empirical research provides support for this process, showing that meta-perceptions play a mediating role in the relationship between others’ appraisals and self-appraisals (Silva et al., 2021; Yue et al., 2021). For instance, reflected appraisals of delinquency from others contribute to an individual’s own view of themselves as delinquent and subsequent criminality (Walters, 2016). Neuroimaging research highlights this tangled nature of self-perceptions and meta-

perceptions, suggesting that individuals recruit social-cognitive brain regions during self-perception (Jankowski et al., 2014; Pfeifer et al., 2009) and that both forms of appraisals engage overlapping core self regions of the default mode network, such as the medial prefrontal cortex (MPFC), the posterior cingulate cortex (PCC), and the inferior parietal lobe (IPL; Delahoy et al., 2022). These findings illuminate meta-perception as a core socio-cognitive process informing self-view.

One advantage to thinking about the role of meta-perception in self-esteem is that it allows us to move away from associative learning mechanisms towards belief-based models, which better capture the uncertainty present in ambiguous social situations. Belief-based models also provide more clinical relevance as they can inform cognitive models of low self-esteem (Low et al., 2021). Additionally, considering the role of meta-perceptive processes across interpersonal theories of self-esteem allows us to explore whether differences in meta-perceptive abilities and tendencies underlie individual differences in self-esteem. We already know there is wide variance in the strategies and capacities individuals possess for mentalizing (for review see: Guazzelli Williamson & Mills, 2022). By illuminating the role mentalizing might play in self-esteem formation, it follows that individual differences in this socio-cognitive process contribute to variance in self-esteem. This also affords the possibility of theory integration – rather than needing to crown a single theory of self-esteem as “correct” or “incorrect”, we can determine what theory applies when and for whom.

Individual Differences in Meta-Perception Processes

In searching for commonalities across divergent theories of self-esteem, we find meta-perception as an implicit process supporting the incorporation of social information into self-view. This opens the possibility that individual differences in meta-perceptual accuracy and tendencies in

part underlie differences in self-esteem. Here we can consider several different dimensions that modulate the relationship between meta-perception and self-esteem: *propensity* (the degree to which one engages in meta-perceptual processes), *accuracy* (the degree to which meta-perceptions correspond with actual appraisals), and *permeability* (the degree to which meta-perceptions inform self-view). The following section will explore how individual differences in meta-perceptions across these domains might inform self-esteem.

Propensity

While most humans tend to be curious about what others think of them, there is variability in the degree to which individuals engage in mentalizing processes in response to this curiosity. Some work demonstrates that individual differences in the *propensity* to mentalize, either overall or in specific contexts, can have downstream effects on self-esteem. This section will review research that explores the relationship between self-esteem and the propensity to engage in meta-perceptual processes to demonstrate how selective disengagement from perspective-taking can support higher self-esteem.

Evidence suggests that avoiding perspective-taking can benefit one's self-esteem. We can consider these findings in the context of the extensive literature on self-protective biases, which show that when faced with negative evaluation from others, individuals tend to engage in positive self-reflection to self-enhance rather than integrate the negative evaluation into their self-view (Anderson & Hanslmayr, 2014; Hepper et al., 2011; Hughes & Beer, 2013; Korn et al., 2012). Individuals with higher self-esteem have increased self-referential processing during negative feedback, suggesting that they are recruiting positive self-schemas to buffer the negative appraisals of others (van Schie et al., 2018). Crucially, this reflects a disengagement

from perspective-taking, as this process is characterized by ignoring negative feedback rather than increasing attention towards positive information about the self from others (Leitner et al., 2014). Moreover, positively biased self-evaluations can be attenuated when a monetary incentive is given for accurate meta-perceptions (Y.-H. Kim et al., 2010) suggesting that recruiting positive self-schemas to reinforce a favorable self-view is more reflective of a propensity to disengage from the appraisals of others rather than an *inability* to do so. Rather than the common adage that high self-esteem rests on “not caring what others think”, it appears that “not *thinking* what others think” can be crucial for maintaining elevated self-esteem. In the face of negative social information, disengaging from perspective-taking to instead bring a positive self-view to mind helps to maintain self-esteem.

Neuroimaging work has identified the *medial prefrontal cortex* (MPFC) as a brain structure central to this process (Hughes & Beer, 2013; Somerville et al., 2010; Van de Groep et al., 2021; Yankouskaya & Sui, 2021), situating the MPFC as a candidate region for the integration of social, self, and reward processing that might compute self-protective responses in the face of negative social feedback (Will et al., 2017). MPFC activation following social feedback can predict individual differences in self-esteem (Eisenberger et al., 2011; Yang et al., 2016) and the MPFC seems to track evaluations from others (Ito et al., 2020) and prompt protective self-reflection in response to negative cues to buffer their debasing effect on self-esteem (Korn et al., 2012; L. Yoon et al., 2018). When the MPFC is disrupted with transcranial magnetic stimulation, we see a correspondent disruption in the self-positivity bias (Duran et al., 2020); such disruption leads individuals to become more likely to attribute positive traits to a peer than themselves (Luber et al., 2012) and predict others will socially accept them at chance level rather than above chance (D. Zhang et al., 2022). Recent work demonstrates that lower self-

esteem is associated with increased correspondence between the neural pattern when an individual reflects on themselves and the pattern when their peers reflect on them; notably, this effect is observed in the MPFC (Stendel et al., 2023). These neural findings highlight that disengaging from social feedback and recruiting an internal, positive self-concept can be essential to maintaining high self-esteem and that the MPFC is a core structure in this process.

If high self-esteem is in part characterized by ignoring negative social cues, we would expect that lower self-esteem might rest on a maladaptive awareness of such cues. Indeed, there is some evidence to suggest that increased perspective-taking is associated with lower self-esteem and a reduction in self-enhancement. Individuals with lower self-esteem have stronger insight into the mental states of others (Talaifar et al., 2021) and seem to engage in more perspective-taking during self-reflection than their high-self-esteem counterparts (Pfeifer et al., 2009). We also see that increased perspective-taking can exert influence on self-esteem in the short term, whereby reflecting on what others think reduces positive bias in self-evaluations (A. Zhou et al., 2013). It is worthwhile to note that mentalizing processes tend to come online in tandem with dips in self-esteem, with the emergence of perspective-taking abilities overlapping with the development of self-awareness (Frith & Frith, 2006; Sahi & Eisenberger, 2021).

Taken together, these findings show that individuals who are more likely to engage in meta-perception tend to have lower self-esteem, and that actually *ignoring* negative social evaluations to instead activate self-protective, positive schemas is vital to preserving high self-esteem. We can contextualize these findings within existing theories of self-esteem by considering the possibility of individual differences in Kinch's *causal model*. While some individuals ignore the appraisals of others with inaccurate, yet positive self-enhancement, others might instead more heavily weigh the path from the appraisals of others to reflected appraisals to

self-appraisals. Alternatively, through the lens of *sociometer theory*, the overweighting of this path might actually be a consequence, rather than a cause, of low self-esteem, where individuals increase their attunement to the social environment in order to adjust their behaviors to increase the probability of social acceptance. In both cases, there seems to be a tradeoff between “turning off” meta-perception to self-enhance vs. heightening social awareness to increase the accuracy of meta-perceptions: Kurt & Paulhus (2008) found that self-enhancement leads to more positive intrapersonal outcomes (such as self-esteem) yet more maladaptive interpersonal outcomes.

Accuracy

Whether ignoring negative social information to maintain self-positivity or incorporating it to reify a negative self-concept, individuals selectively use social information to preserve their self-view (Yoon et al., 2018). This opens the possibility that biases, both in self-view and in the social information one chooses to attend to and integrate — can shift meta-perceptions to influence self-esteem. This section will examine *accuracy* as another dimension that we can explore when considering the relationship between meta-perceptions and self-esteem. Here, “accuracy” refers to high levels of agreement between meta-perceptions and the actual perceptions of others, also known as “meta-accuracy.” *How closely do our meta-perceptions match the actual perceptions of others, and how does this inform self-view?*

Although to variable degrees, most humans devote significant energy towards observing and interpreting the behaviors of others. Yet despite these efforts, our interpretations about the inner worlds of others are often incorrect (Byron & Landis, 2020; Kenny, 1994). All of us have experienced select instances when we fail to accurately infer what another is thinking, but research also demonstrates that these inaccuracies can cluster as part of broader tendencies to

cast our meta-perceptions in an overly positive or negative light. One of these tendencies with robust empirical support is the *self-enhancement bias* (Krueger, 1998), or the pervasive tendency for humans to overestimate their own abilities and positive qualities. Self-enhancing biases span across a wide range of domains and frequently emerge within interpersonal contexts. Several studies demonstrate that people often make overly optimistic predictions when trying to determine whether peers or strangers will accept them (Cao et al., 2015; Hepper et al., 2011; Loeb et al., 2016; Loeb et al., 2018; Petereit et al., 2022; Van Der Molen et al., 2014). Individuals with higher self-esteem believe they are valued by others (Bleckmann et al., 2022); critically, this effect is due to biased meta-perceptions rather than any demonstrated accuracy that they are actually more valued than others around them (Back et al., 2009).

The tendency to hold a positive self-view in disagreement with the appraisals of others is tied to motivation and reward processing. Individuals who demonstrate stronger frontal-striatal functional connectivity in response to negative social evaluation have significantly higher state and trait self-esteem (Shany et al., 2022; Yoon et al., 2018), suggesting one avenue by which self-enhancement might restore self-worth in the face of social information that threatens self-positivity. Several studies demonstrate that higher self-esteem is associated with stronger functional and structural connectivity between brain regions involved in self-processing and reward-processing (Chavez & Heatherton, 2015; Northoff & Hayes, 2011). Taken alongside behavioral findings, it seems that when faced with negative perceptions from others, recruiting inaccurate yet positive self-perceptions and meta-perceptions can restore self-esteem.

While some research shows that high self-other agreement in personality predicts higher self-esteem (Kim et al., 2019; Kim et al., 2010; Luan et al., 2018), no known existing work links higher meta-accuracy to self-esteem. In fact, broader meta-accuracy research shows that more

psychologically adjusted individuals make less accurate meta-perceptions (Carlson, 2016; He & Côté, 2019; Mosch & Borkenau, 2016), providing more support that positive meta-perceptual biases are a form of self-enhancement supporting heightened self-esteem.

Despite robust evidence supporting the self-enhancement bias, it is crucial to note that not everyone self-enhances. Unsurprisingly, lower self-esteem is associated with a tendency to self-*efface* rather than enhance, and as such is often characterized by negatively biased meta-perceptions (Gruenenfelder-Steiger et al., 2016; Moritz & Roberts, 2020; Rouault et al., 2021; Wagner et al., 2018). These meta-perceptions are more strongly informed by self-perceptions than the actual appraisals of others (Bleckmann et al., 2022; Hu et al., 2014; Kaplan et al., 2009), suggesting that negatively biased beliefs about ourselves can debase our meta-perceptions. Moreover, meta-perceptions predict self-esteem more than the actual appraisals of others (Gruenenfelder-Steiger et al., 2016; Wagner et al., 2018). Taken together, these studies highlight that when it comes to our self-esteem, our perceptions of whether others approve of us matter more than their actual opinions about us. This body of work shows that negative meta-perceptions — even when entirely divorced from the actual appraisals of others — can be a feature of low self-esteem.

It is important to note the discrepancy between these findings that suggest that low self-esteem is characterized by negatively biased meta-perceptions and the earlier described work showing that low self-esteem is characterized by more accurate meta-perceptions (e.g. Talaifar et al., 2021). As of now, it remains unclear whether low self-esteem is better characterized by negatively biased meta-perceptions or accurate meta-perceptions, or whether a non-linear relationship exists whereby we see different associations emerge at different thresholds of self-esteem.

Meta-perceptive biases are also context-dependent, with the target and content of the meta-perception influencing its effect on self-esteem. Some work shows increased reward processing in the brain in response to positive feedback but only when this feedback comes from well-liked individuals (Hughes et al., 2018). Meta-perceptions are better remembered when they come from individuals that we feel closer to (Yue et al., 2020). Meta-perceptions about traits that are more evaluative, morally charged, or central to our self-concept might be more relevant to self-esteem (D'Argembeau et al., 2012; Elder et al., 2021; John & Robins, 1993; Strohming & Nichols, 2014). There also might be within-person variability in meta-perceptive biases that are situationally dependent; we might weigh out different goals and motivations to decide whether to self-enhance or self-verify (Lynch & vanDellen, 2020). Ultimately, considering who is evaluating us and what we are being evaluated on might be important when considering the relationship between biased meta-perceptions and self-esteem.

While contextual factors might shift these effects, this overall pattern of results suggests that self-esteem can be both inflated by positively biased meta-perceptions and diminished by negatively biased ones. This effect appears particularly salient when we are confronted with social information that threatens our self-worth. In such scenarios, engaging in positive, inaccurate self-referential processing, rather than attending to the negative value cues of others, is an essential response to restore or maintain our self-esteem. Still, while individuals do demonstrate consistent biases towards positive and negative meta-perceptions, more work is needed to better understand how these biases operate across different contexts and how accurate meta-perceptions relate to self-esteem.

Permeability

Another dimension we can explore when considering the relationship between meta-perceptions and self-esteem is examining the dimension of *permeability*: to what degree do an individual's self- and meta-perceptions fluctuate in response to the perceptions of others, or how vulnerable is one's self-concept to external influence? Put differently, how permeable is the boundary separating our intrinsically informed self-perceptions from those informed by others? The present literature features several theoretical accounts that propose that low self-esteem is characterized by both high and low permeability.

According to theories that propose a link between low self-esteem and a highly permeable self-concept, low self-esteem is theorized to emerge from an overly context-dependent self, whereby the more an individual uses the perceptions of others to inform their self-concept, the lower their self-esteem. This view is based in the theory of *self-concept clarity*, defined as “the extent to which the contents of a person's self-concept are clearly and confidently defined, internally consistent, and temporally stable,” (Campbell et al., 1996, p. 141). Brockner's Plasticity Hypothesis (Brockner, 1988) links this concept to self-esteem, positing that those with lower self-esteem are more sensitive to social feedback. Some research suggests that those with lower self-concept clarity show increased awareness of the social environment (Hogg et al., 2007), turning to the interpersonal realm to create a more interdependent self-construct (Cabooter et al., 2016). In other words, when an individual does not have a clear sense of who they are, they increasingly rely on interpersonal information and social evaluations to inform their self-concept. This dovetails with work on the relational self, which suggests that holding a highly context-dependent self-construal is reciprocally linked to a strong need for belonging,

which together lead to downstream cognitive, motivational, and emotional shifts that increase one's sensitivity to the thoughts and feelings of others (Suh, 2007).

Research demonstrates that this tendency to attune to the social world and subsequently integrate social information into the self-concept can impair general psychological well-being (Butler et al., 1994; Harter et al., 1996; Kernis & Goldman, 2003) and is a feature of lower self-esteem. For instance, while self-esteem is generally responsive to value cues from others (Weisbuch et al., 2009), these value cues seem to exert more influence on the self-view of individuals with lower self-esteem (Rodman et al., 2017; Tice & Wallace, 2003). Those with lower self-esteem show increased sensitivity to signs of social rejection (H. Li et al., 2012) and social approval, and their meta-perceptions about their social approval exert greater influence on their self-esteem than their actual social rank (Low et al., 2021). Low self-esteem is often associated with high rejection sensitivity (Zhou et al., 2020), and these expectations of rejection guide how individuals participate in and learn from social feedback (Kube & Rozenkrantz, 2021). Those who have heightened rejection sensitivity show increased activation in the mentalizing brain network when faced with social rejection (Sahi & Eisenberger, 2021), signifying a turn to the social world in response to negative social appraisals, rather than the recruitment of positive, internal self-schemas that we saw to be characteristic of individuals with high self-esteem. Altogether, this body of work suggests that defining oneself on internal standards, rather than on external social information, is critical for preserving a positive self-view — or in other words, lower self-concept permeability maintains higher self-esteem.

Another line of work complicates this picture by suggesting that individuals with lower self-esteem can have *less* socially permeable self-views. On this view, these individuals approach social situations with strong priors and give too much precision to prior beliefs, leading to the

persistence of negative self-view even in the face of disconfirming information (Friston et al., 2014; Paulus et al., 2019). This appears to be valence dependent, such that those with lower self-esteem are less likely to update their self-views from positive peer appraisals (Elder et al., 2021; Josephs et al., 2003; Will et al., 2020). While “healthy” individuals are more likely to change their self-view in response to better-than-expected social feedback than worse-than-expected social feedback (Koban et al., 2017; Korn et al., 2012; Sharot et al., 2011), individuals with low self-esteem demonstrate the opposite learning profile, showing a reduced tendency to update their self-views following positive feedback (Shany et al., 2022). This tendency can in part be attributed to differences in reward learning due to attenuated reward response in the ventral striatum in response to positive social feedback (Shany et al., 2022). In sum, this set of literature shows that having an inflexible, negative self-concept that is resistant to influence from positive social feedback is also characteristic of low self-esteem.

Some research has attempted to reconcile the conflicting findings that low self-esteem is both characterized by a stable, negative self-concept as well as an unstable self-concept. Older findings highlight a *Simpson's paradox*, which helps explain the crux of this contradiction. While overall, a positive, linear relationship is observed between self-esteem and self-concept clarity, at low levels of self-esteem the direction of the effect is reversed such that instability in self-concept corresponds with higher self-esteem (Kernis et al., 1989). For individuals with lower self-esteem, lower self-concept clarity might at least afford the occasional experience of positive self-representations that can contribute to an overall more positive self-concept. More recent work also reconciles conflicting findings about the relationship between self-concept clarity and self-esteem by examining the valence of appraisals. Individuals with low self-esteem show deficits in reinforcement learning such that they are less likely to update their own

perception of their social worth in response to positive social information (Will et al., 2020), while also demonstrating a bias that overweighs negative social information that is congruent with their negative self-concept (Kube & Rozenkrantz, 2021). In other words, individuals with lower self-esteem demonstrate quick, volatile shifts in their self-view in response to social feedback yet are slow to update negative social expectations (Will et al., 2020). This contrasts with the social learning parameters of individuals with healthy levels of self-esteem, who integrate positive self-congruent information and ignore negative, incongruent information, to maintain an overall positive self-view (Korn et al., 2012). Either having too strong of a data-driven or theory-driven self-concept can pose a risk to self-esteem. Together, these findings suggest that low self-esteem is characterized by a self-concept that is both permeable and resistant to social influence, critically depending on the valence of the social appraisal in question.

In synthesizing existing literature across different dimensions of meta-perception (accuracy, propensity, and permeability), we find that engaging in positive, inaccurate self-referential processing, rather than forming meta-perceptions about what others think of us, is an essential response to restore or maintain high self-esteem. Additionally, research demonstrates that lower self-esteem is supported by both negatively biased meta-perceptions as well as fewer positively biased meta-perceptions. Finally, we see that low self-esteem is characterized by altered social learning such that an individual with lower self-esteem is highly sensitive to negative perceptions of others, yet can fail to learn from positive peer perceptions. Since low self-esteem is a vulnerability factor for the later development of internalizing disorders, it is crucial that we continue to refine our understanding of socio-cognitive processes that support

Table 2.1. Meta-perceptive Accuracy, Permeability, and Propensity: Description and Examples

Tendency	Description	Example	Adaptive Functions	Maladaptive Functions
Propensity	The degree to which an individual engages in meta-perceptive processes.	Mary frequently thinks about what others are thinking of her, thus having a higher <i>propensity</i> towards meta-perception than Olga, who rarely considers what others think about her.	Perspective-taking allows us to consider the viewpoints of others and adjust our behavior accordingly to achieve interpersonal goals.	Frequently considering the potentially negative perspectives of others can degrade self-esteem and make it more difficult to recruit internal, positive self-representations.
Accuracy	The degree to which meta-perceptions correspond with the actual appraisals of others.	Mary believes that her teacher thinks she is well-behaved, and her teacher does hold this perception, thus Mary's meta-perception is <i>accurate</i> .	Holding more accurate meta-perceptions allows us to adjust our own behavior to increase the likelihood of positive interpersonal outcomes.	Being more aware of the actual perceptions others hold can disrupt the self-enhancement processes that support self-esteem.
Permeability	The degree to which the perceptions of others inform self-view.	Both Mary and Olga might hold the same meta-perception that others think they are unintelligent. If Mary had higher <i>permeability</i> , this meta-perception would influence her self-view more than it would for Olga.	Our identities are in part formed by the perceptions of others; having a permeable self-concept allows us to build an identity rooted in our social interactions with others.	Since others typically only see a portion of who we are, often in certain contexts, relying too heavily on meta-perceptions to inform self-view might prevent a multidimensional and stable self-concept.

transdiagnostic risk and might be effective targets for prevention and intervention, like meta-perceptual processes.

Internalizing Symptoms as a “Perfect Storm” of Vulnerability

Although self-esteem is implicated across many mental disorders, it is a particularly relevant clinical construct when considering *internalizing* disorders (Keane et al., 2017; Ngo et al., 2020), a descriptive label capturing diagnoses that feature persistent negative emotions directed inward, rather than outwardly expressed, such as in unipolar mood and anxiety disorders (Eaton et al., 2013). Moreover, internalizing disorders are characterized by concurrent disruptions in self-referential processing, social cognition, and reward processing that might interact to create a ‘perfect storm’ for maladaptive mentalizing tendencies.

Aberrant self-referential processing is observed across internalizing disorders, which are frequently characterized by increased self-focus and negative beliefs about the self (Ginot, 2012; McEvoy et al., 2019; Mor & Winquist, 2002). Various studies demonstrate elevated self-referential processing in depression (Kuiper & Derry, 1982; Lemogne et al., 2012; Y. Li et al., 2017; Nejad et al., 2013; Northoff & Tumati, 2019), social anxiety disorder (Abraham et al., 2013; Lin et al., 2018; Yoon et al., 2019), and generalized anxiety disorder (Cui et al., 2022), particularly *negative* self-referential processing such as rumination and self-criticism (Dinulescu et al., 2021; Ke et al., 2020; Kircanski et al., 2015). Neuroimaging research has demonstrated aberrant activity in brain regions and networks involved in self-processing (Lemogne et al., 2009; Lin et al., 2018; Nejad et al., 2013; Northoff & Tumati, 2019), lending further support that elevated negative self-focus is one factor that contributes to internalizing symptomatology.

Internalizing disorders are also characterized by disruptions to social processing. Social withdrawal, either due to fear or anhedonia, is common across internalizing disorders (Rappaport & Barch, 2020), along with reductions in social support (Rueger et al., 2010), increased rejection sensitivity (Ayduk, et al., 2001; Heeren et al., 2017; Harrewijn et al., 2017; Hsu et al., 2015; Jankowski et al., 2018; Kumar et al., 2017; Silk et al., 2014), and socio-cognitive deficits (Dainer-Best et al., 2017; Gotlib & Joormann, 2010).

Internalizing disorders also feature broad disruptions in reward processing (Kujawa & Burkhouse, 2017), such as hypersensitivity to punishment/negative stimuli (Eshel & Roiser, 2010; Ng et al., 2019), hyposensitivity to reward/positive stimuli (Bylsma et al., 2008; Eshel & Roiser, 2010; Feeser et al., 2013; Moran et al., 2012), and reduced motivation to pursue rewarding stimuli (Bryant et al., 2017). Notably, these global differences in reward processing also lead to aberrant reward processing of self-and socially- relevant information, and indeed we see that internalizing symptoms are linked to attenuated reward response to self-relevant stimuli (Northoff & Hayes, 2011; Quevedo et al., 2017) and positive social feedback (Davey et al., 2011; Z. He et al., 2019; Zhang et al., 2020; Zhang et al., 2017).

While the role of self-, social-, and reward-processes in internalizing disorders are often studied independently, emerging work has attempted to integrate research across these areas (e.g. Luyten & Fonagy, 2018; Pfeifer & Allen, 2021). Recent research suggests that this integration might be crucial when investigating internalizing disorders; for instance, Hobbs et al. (2021) show that while depression symptoms cannot be predicted by independent alterations in self-referential, emotion, or reward processing, they *are* strongly predicted by performance on tasks that require the coordination of these three domains.

As such, there have been calls for increased attention to processes that lie at the intersection of self, social, and reward processes, such as mentalizing (Luyten & Fonagy, 2018). Such work proposes that convergence between reward processing, stress regulation, and the emerging capacity for mentalizing can create a feedback loop that leads to the maintenance of internalizing symptoms (Luyten & Fonagy, 2018). On this view, attenuated sensitivity to positive feedback reduces one's ability to learn from positive feedback about the self. Similarly, increased sensitivity to punishment increases the salience of negative or ambiguous feedback, reifying negative self-schemas. These maladaptive self-schemas then induce cascading changes in attention, reward, and social processes which reinforce those very schemas, trapping one in a vicious circle (Luyten & Fonagy, 2018). Notably, the fact that these very socio-cognitive processes come online during adolescence might in part explain why this developmental period marks a massive increase in risk for internalizing disorders (Pfeifer & Allen, 2021).

The notion that negative self-schemas and aberrant social information processing can mutually reinforce one another is evident in the self-esteem literature, which demonstrates that low self-esteem is characterized by attenuated positivity bias, altered reactions to social information, and self-schemas that are highly sensitive to negative feedback yet impervious to positive feedback. In drawing parallels between the social psychology literature on meta-perception and self-esteem and the more clinically oriented mentalizing literature, we can explore one avenue by which low self-esteem might confer vulnerability to internalizing disorders.

Summary of Mentalizing Research across Internalizing Symptoms

Although mentalizing has been identified as a core process that might support the maintenance of internalizing disorders (Luyten & Fonagy, 2018), limited and conflicting

research findings have muddled its precise relationship to internalizing symptomatology. Generally, broad mentalizing deficits have been observed in individuals with internalizing disorders. Impairments in theory of mind are demonstrated in both generalized anxiety (Berg et al., 2021) and depression (Harkness et al., 2011; Lee et al., 2005; Nejati et al., 2012; Nestor et al., 2022). In depression, these deficits persist across a variety of tasks that capture different facets of mentalizing, including emotion decoding (Lee et al., 2005) and mental state reasoning (Cusi et al., 2013). Importantly, these deficits remain even when controlling for baseline cognitive dysfunction that can occur with depressed mood (Luyten & Fonagy, 2018). Some work has shown that more sophisticated mentalizing abilities are positively associated with psychological well-being, suggesting that accurate mentalizing might serve a protective function against psychopathology (Ballespí et al., 2021).

On the other hand, a significant body of work shows the opposite, instead finding *enhanced* mentalizing abilities in internalizing disorders like depression (Harkness et al., 2010; Lee et al., 2005; Mellick & Sharp, 2016) and social anxiety disorder (Berg et al., 2021; Tibi-Elhanany & Shamay-Tsoory, 2011). Causal evidence supports this view by demonstrating that inducing sad mood increases mentalizing accuracy (Converse et al., 2008) and vice versa for positive mood (Harkness et al., 2010). These findings are consistent with research on *depressive realism*, which shows that the negative cognitions observed in depression reflect a more accurate assessment of the world that lacks the positive bias that supports psychological well-being (Alloy & Abramson, 1988; Moore & Fresco, 2012).

How can we reconcile the conflicting findings that internalizing disorders are associated with both enhancements and deficits in mentalizing? Some propose a need to differentiate between different forms of mentalizing (e.g. decoding vs. reasoning, Maleki et al., 2020;

emotional vs. cognitive, Chepenik et al., 2007) or disorder subtype (for instance, enhanced theory of mind might only be present in depression cases characterized by social rejection; Hanna, 2015). Another explanation might be that the relationship between internalizing symptoms and mentalizing abilities is non-linear, whereby enhanced abilities are observed at low levels of symptomatology but not maintained at more elevated levels of distress. This is corroborated by findings showing that enhanced mentalizing skills are more consistently observed in individuals with sub-clinical internalizing symptoms than those who meet threshold for a disorder (Bodenhausen et al., 1994; Lee et al., 2005; Schreier et al., 2013), or who have a history of major depression but are in remission (Harkness et al., 2010). Some work shows a quadratic relationship, where both low and high levels of cognitive empathy are associated with depression (Tully et al., 2016). Ultimately, more work is needed to understand how internalizing disorders can feature both increased social sensitivity *and* broad perspective-taking deficits.

Meta-Perception in Internalizing Disorders

Mentalizing underlies our ability to form meta-perceptions, which by nature requires us to infer what others are thinking about us. When we examine how mentalizing tendencies influence meta-perceptions in internalizing disorders, we see the same paradox emerge, whereby individuals demonstrate both enhanced and diminished meta-perceptive abilities. These abilities often follow a pattern of “over-mentalizing”, whereby individuals with internalizing symptoms more readily form meta-perceptions, but these meta-perceptions tend to be negatively biased (O’Connor et al., 2007; Washburn et al., 2016). Over-mentalizing can be understood as an adaptive response to actual or potential social disapproval whereby individuals are more motivated to understand what others are thinking about them in order to restore social status and

minimize social risk (Allen & Badcock, 2003; Andrews & Thomson Jr., 2009; Gilbert & Miles, 2000; Weary & Edwards, 1994). Social exclusion induces low-level perceptual and behavioral shifts that increase sensitivity to social cues of rejection (Kawamoto et al., 2014). Given that social rejection is highly associated with risk for internalizing disorders (Kendler et al., 2003; Slavich et al., 2010) paired with the fact that social deficits might increase the motivation to perspective-take, it follows that hypersensitivity to the mental states of others would be an effective meta-perceptive strategy to protect against further rejection, even at the cost of psychological well-being. For example, strong expectations of being rejected buffer against negative affect following social rejection in depressed individuals (Caouette & Guyer, 2016), representing a form of *defensive pessimism* (Norem & Cantor, 1986). For individuals who expect to be rejected (either due to cognitive biases or an actual history of social rejection), adopting a more conservative meta-perception strategy to minimize social risk might seem adaptive.

This more conservative meta-perception strategy can both cause and reinforce negative self-beliefs, which are already a core element of internalizing disorders (Clark & Wells, 1995; Hofmann, 2007; Moscovitch, 2009). Naturally, such beliefs influence meta-perceptions, where individuals with internalizing disorders frequently report negative beliefs about what others might think about them. Individuals with depression underestimate their social rank, level of affiliation, and how much others like them (Bao et al., 2022; Bone et al., 2021; Zabag, 2018) and individuals with social anxiety underestimate how positively others see them (Zabag, 2018). In addition to negatively biased meta-perceptions, individuals with internalizing disorders frequently lack the positive bias in meta-perceptions typical of healthy controls (Goldin et al., 2009). Overall, we see that internalizing disorders feature meta-perceptions that are both more negatively and less positively biased.

These biases in meta-perceptions can contribute to the expectation of negative social feedback and subsequently reinforce negative self-views in a cyclical fashion (See Figure 2.1). This cycle unfolds via *motivated perception*, whereby strong priors influence lower-level perception to shift how novel information is interpreted, ultimately selectively favoring information that confirms existing higher-order self-schemas (Kube & Rozenkrantz, 2021). Aberrant responses to social feedback are frequently found amongst internalizing disorders in line with models of motivated perception (Hsu & Jarcho, 2021; Rappaport & Barch, 2020). Individuals with internalizing disorders show attentional biases towards negative social cues (Schultz & Heimberg, 2008) and interpret ambiguous social feedback as more negatively biased (Everaert et al., 2017) and less positively biased (Amir et al., 2012). These biases lead to an over-weighting of negative feedback, where individuals with both social anxiety and depression have higher learning rates for negative traits about themselves and more readily update their self-beliefs in response to negative feedback (Koban et al., 2017; Korn et al., 2012). Moreover, these individuals show slower learning rates from positive social feedback, failing to update meta-perceptions when faced with new, positively-valenced social information that disconfirms their self-views (Everaert et al., 2017; Kube & Rozenkrantz, 2021; Liknaitzky et al., 2017). Together, increased precision given to prior beliefs and reduced weight of disconfirming information can operate reciprocally to support the persistence of meta-perceptions aligned with the core beliefs of the disorder (Friston et al., 2014; Kube & Rozenkrantz, 2021; Paulus et al., 2019).

Given that low self-esteem is a large risk factor for the later development of internalizing psychopathology, bridging these literatures can help us understand how altered mentalizing processes occurring in low self-esteem connote transdiagnostic risk for internalizing disorders. The summarized findings above suggest that low self-esteem can serve as an entry point to a

maladaptive, positive feedback loop where negative meta-perceptions, attentional biases, and aberrant social learning mutually reinforce one another to worsen symptoms (See Figure 1). Dips in self-esteem can be adaptive in that they attune us to the social world to help us restore social inclusion or support key developmental tasks such as the rapid learning of social norms critical in adolescence. However, more exaggerated drops might shift meta-perceptions to be excessively negative, rather than more accurate, leading to downstream changes in attention, behavior, and learning parameters that all mutually reinforce one another. For instance, the cycle might unfold as follows: an individual might become more sensitive to negative social cues, which leads to increased social withdrawal in order to reduce social risk. This reduces potential exposure to evidence that contradicts their negative self-perceptions, and they begin to overemphasize any feedback that confirms their negative beliefs. This then strengthens their negative priors, which are consistent with the core beliefs of the disorder, reinforcing the maladaptive attentional biases and social behaviors.

This cycle might explain why we see more accurate meta-perceptions at subclinical levels of psychopathology (when individuals are lacking the self-enhancement and thus seeing themselves more accurately), yet more inaccurate meta-perceptions once in the throes of a disorder (when negative self-views become more entrenched and negatively biased and thus distanced from reality). Here, accurate mentalizing can be viewed as a "risky strength" (Tone & Tully, 2014) offering some interpersonal benefits while conferring intrapersonal risk. Ultimately, examining the differences in meta-perception tendencies across differing degrees of psychopathology provides clues as to how low self-esteem increases risk for internalizing disorders, especially considering alterations to self, social, and reward processing in a

maladaptive cycle. Also, it allows us to explore how intervening on different parts of this cycle might confer therapeutic benefits.

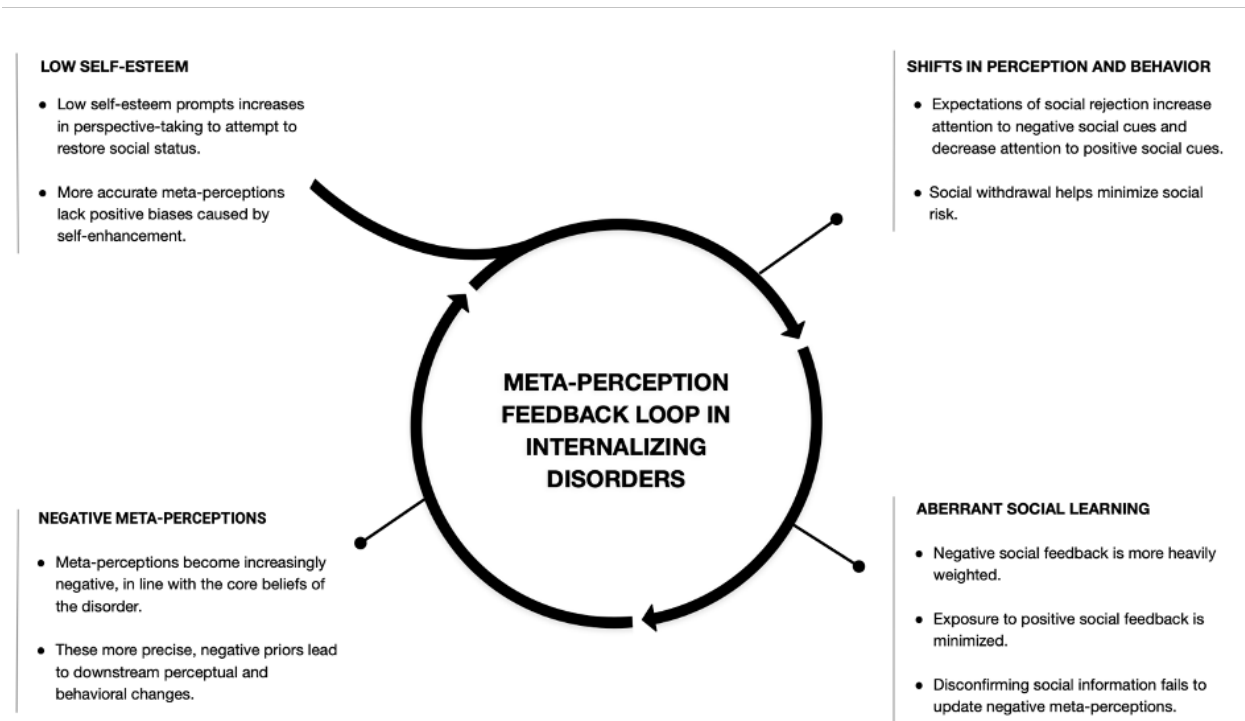


Figure 2.1. Meta-perception feedback loop in internalizing disorders. This figure depicts how self-referential processing, social cognition & behavior, and changes in reinforcement learning interact to maintain or worsen internalizing symptoms. Critically, we highlight how low self-esteem might serve as an entry point to this cycle, initially prompting an individual to form more frequent and accurate meta-perceptions to restore social status, but ultimately leading to downstream perceptual, behavioral, and learning changes that reinforce excessively negative meta-perceptions.

Therapeutic Implications and Treatment Targets

Given the degree to which negative meta-perceptions contribute to internalizing symptoms, it is no surprise that they are a common target of intervention in psychotherapy. Both

cognitive-behavioral and psychodynamic approaches have increasingly begun to address deficits in mentalizing in internalizing disorders (Luyten & Fonagy, 2018; Watkins & Teasdale, 2004).

Cognitive strategies for treating internalizing disorders target maladaptive thoughts, inferences, and interpretations that contribute to symptom maintenance and reinforce negative self-schemas (Beck & Weishaar, 1989). Such approaches might have individuals restructure unhelpful meta-perceptions by generating more balanced and accurate meta-perceptions. It is presently unclear whether challenging underlying self-perceptions (e.g. “I am stupid”) is more effective than challenging meta-perceptions (“My friends think I am stupid”). On one hand, self-perceptions strongly inform meta-perceptions (Bleckmann et al., 2022) which might make these a more effective target, but on the other hand, meta-perceptions that disconfirm negative self-beliefs might be more amenable to change because they do not require an individual to believe something positive about themselves. Targeting meta-perceptions might also help shift social information processing to support the integration of evidence that disconfirms negative self-views. Moreover, meta-perceptions might be a particularly salient target during developmental periods characterized by strong social salience, such as adolescence. Additional work is needed to identify how to best address distorted meta-perceptions that are contributing to symptoms.

Other work proposes additional cognitive mechanisms that might be effective psychotherapeutic targets. For instance, clients might benefit from increasing positive self-views by rehearsing the activation of positive evidence or self-compassion in response to social feedback (Arch et al., 2018; Hopkins et al., 2019; Kube et al., 2019). We know that self-enhancement is causally related to mood (Duran et al., 2021) and that Cognitive-Behavioral Therapy can increase positive self-views (Goldin et al., 2013), suggesting this to be one fruitful area for therapeutic benefit. Other work targets attentional mechanisms, using attentional re-

training to shift attention to positive social cues (Taylor et al., 2011) or to reduce self-focused attention (Lin et al., 2018), which can help clients disengage from the negative, self-referential processing that prevents them from integrating novel, positive information into self-schemas (Erle et al., 2019). Finally, some recent work suggests that increasing meta-cognitive awareness of negative self and social biases by delivering psychoeducation about such biases might be an effective intervention (Hopkins et al., 2021).

Behavioral treatment targets can also indirectly target meta-perceptions. For example, exposures and behavioral experiments are employed to help clients encounter disconfirming information that would challenge unhelpful beliefs. Here, the meta-perception “the cashier will watch me fumble with my change and think I am foolish” might be challenged by having a client attempt to purchase something and either have an encounter with the cashier that disconfirms this belief or learn that they are still able to approach the feared situation even in the face of the maladaptive meta-perception. Behavioral approaches can also help clients encounter positive reinforcers in their environment, like finding groups and social settings where the client is accepted, in turn improving self-image (Miers, 2021). Another general behavioral approach, *behavioral activation* leads to increased positive self-referential processing (Kircanski et al., 2013), suggesting that even broad behavioral approaches might also target the effect maladaptive meta-perceptions might have on internalizing symptoms.

Psychodynamic modalities also involve the modification of mentalizing, either implicitly through a corrective relationship with the clinician (e.g. Huang et al., 2016) or more directly by targeting meta-perception processes themselves, such as with *Mentalization-Based Therapy* (MBT; Daubney & Bateman, 2015). This psychodynamic approach encourages the client to feel understood, or *mentalized*, which leads to increased curiosity about the mental state of the

therapist, ultimately improving mentalizing abilities that increase the client's capacity to learn from positive social influences in their environment (Daubney & Bateman, 2015). Critically, improved mentalizing abilities provide the foundation for individuals to benefit from novel and reparative social experiences (Fonagy & Bateman, 2006). Most of the research on MBT is focused on borderline personality disorder and the evidence base demonstrates positive clinical outcomes in this population (Malda-Castillo et al., 2019). While there is some emerging work supporting its application in internalizing disorders (Bressi et al., 2017; Markowitz et al., 2019; Vogt & Norman, 2019) findings are mixed (Malda-Castillo et al., 2019), possibly because the precise relationship between mentalizing abilities and internalizing symptoms remains unclear. This points to a need to better understand the nuanced relationship between mentalizing and psychopathology as treatments continue to develop in this area (Byrne et al., 2020).

Psychopharmacological treatments also seem to target the “perfect storm” of aberrant social, self, and reward processing that gives rise to maladaptive meta-perceptions. Short-term treatment with selective serotonin reuptake inhibitors (SSRIs) is associated with increased prosociality, positive affective bias, and positive learning from social evaluation (Hobbs et al., 2021) as well as reduced self-referential processing (Komulainen et al., 2018), highlighting these processes as important mechanisms contributing to the maintenance of internalizing symptoms.

Given that shifts in mentalizing processes occur in step with low self-esteem, which renders individuals more vulnerable to internalizing disorders (Orth et al., 2009), it is crucial that future research examine how targeting mentalizing in individuals with lower self-esteem might help prevent or attenuate the development of internalizing symptoms. Indeed, recent randomized-control trials demonstrate the benefits of targeting self-esteem as a cross-diagnostic factor (Zabel et al., 2022) that can improve the quality of life (Martinsen et al., 2021). Future

research is needed to explore how targeting processes core to low self-esteem, such as disrupted mentalizing, might serve as a preventative and transdiagnostic intervention for internalizing disorders.

Conclusions and Future Directions

Aberrant mentalizing, specifically meta-perception (i.e. the thoughts we believe others have about us), is theoretically and empirically related to self-esteem, although the exact nature of this relationship is unclear. We see parallel processes emerge in the literature on internalizing disorders, which we would expect given that meta-perception relies on the coordination of self, social, and reward processes that exhibit disruptions in individuals with internalizing symptoms. Since low self-esteem is a vulnerability factor for later development of internalizing disorders (Orth & Robins, 2022), it is crucial that we continue to refine our understanding of socio-cognitive processes that support transdiagnostic risk and might be effective targets for prevention and intervention, like mentalizing.

Limitations in current research paradigms muddy our understanding of the relationship between psychopathology and mentalizing. Studies use a variety of paradigms that seem to capture different facets of mentalizing (Guazzelli Williamson & Mills, 2022) and often have low levels of convergent validity (Weinstein et al., 2022). Many tasks also imply meta-perceptive processes without explicating them, and thus are siphoned into a different literature, such as studies that make use of the Trier Social Test Stress or Cyberball Task, the latter of which activates mentalizing networks in the brain (Meulen et al., 2016). Across these paradigms, research tends to focus on social exclusion, which is only one form of social feedback (Rappaport & Barch, 2020). Since we see blunted responses to positive stimuli in internalizing

disorders, exploring dysfunctional reactions to positive stimuli, such as social acceptance and inclusion, will also be critical (Hsu & Jarcho, 2021).

Different mentalizing studies attempt to draw comparisons across a variety of relationship types, which also might explain conflicting findings. This work recruits individuals, dyads, and groups at all different levels of acquaintance, from first impressions (i.e. “zero-acquaintance”) to long-term relationships. Some work indicates that mentalizing properties differ across different relationships and levels of acquaintance (e.g. Letzring & Funder, 2018), suggesting that we might see different relationships between meta-perception and psychological well-being when considering relationship factors.

Importantly, our interpretations are constrained by limitations and biases within the existing literature. While conceptions of self are culturally grounded, much of the current literature is conducted in a W.E.I.R.D. (Henrich, 2010) context, limiting the generalizability of findings. The sociocultural identities of the authors of this review also likely constrain the nature of the questions asked. Furthermore, while the present view focuses on meta-perceptions in an interpersonal context, some propose that we consider meta-perception within a socio-ecological framework (Batchelder & Hagan, 2022). Here, “perceptions of one’s mattering and deservingness of equity and psychological, social, and material resources” (Batchelder & Hagan, 2022) – which can be considered a meta-perception about a generalized collective – can significantly contribute to self-worth and the development of psychopathology. While this review has focused on the interpersonal sphere, future work can draw on the rich sociological literature in this area to consider how meta-perceptions at sociocultural and structural levels might inform self-view.

Given these limitations in research on mentalizing and internalizing disorders, significant work is needed to clarify how to best modify this process in clinical work. For instance, do we aim to change maladaptive meta-perceptions, knowing that they influence self-views, or is it more effective to directly target negative, self-relevant cognitions? Is it more useful to directly modify maladaptive meta-perceptions, or instead shift aberrant social learning that might lead to downstream cognitive changes? Should clinicians strive to help clients form accurate meta-perceptions or positive meta-perceptions? While research frequently focuses on the benefits or drawbacks of accurate meta-perception, self-esteem might rest on integrating more abstract value cues about the self, and there is evidence for a distinction between “knowing” and “valuative” processes that might differentially influence self-worth (Wessels et al., 2020). Before even being able to answer such questions, we also might need to improve our ability to quantify self-views, not only for guiding treatment but also to help assess changes in therapy beyond symptom reduction (Taylor & Montgomery, 2007).

In light of the existing literature reviewed, this paper highlights the critical role of mentalizing processes in shaping self-esteem, and underscores the relevance of these processes for understanding and treating transdiagnostic constructs such as low self-esteem in the context of internalizing disorders. By shedding light on how individuals selectively incorporate or ignore social appraisals through different meta-perception strategies, this work offers insights into how mentalizing tendencies can both maintain and exacerbate symptomatology and suggests avenues for linking work from social and clinical psychology for future research and clinical interventions.

CHAPTER III

SELF-ESTEEM MODULATES THE BRAIN-TO-BRAIN SIMILARITY OF THE REPRESENTATION OF THE SELF IN THE BRAINS OF OTHERS

This work is in revision and was co-authored with T. Guthrie, V. Guazzelli Williamson, and R.

Chavez. I was the lead author of this publication and responsible for the writing of the manuscript with edits provided by co-authors. I led the study design, methods, and analyses, with input and assistance from co-authors.

Introduction

Self-esteem – an individual’s subjective evaluation of their own worth – is believed to develop alongside social factors such as close relationships and social support. The connection between self-esteem and social relationships has been theorized as early as Cooley’s looking-glass self (Cooley, 1902), the idea that our self-perceptions are informed by our perceptions of how others view us. Indeed, even competing models of self-esteem highlight interpersonal perception as a core process supporting this reciprocal development of self-esteem and social relationships. For instance, Swann’s self-verification theory (Swann & Read, 1981) suggests that self-esteem is not dependent on being perceived positively by our peers, but rather by being perceived accurately. In contrast, self-enhancement theory (Krueger, 1998) states that self-esteem is maintained by an intrinsic motivation to view oneself in a positive light, even at risk of ignoring or minimizing negative peer-perceptions (Anderson & Hanslmayr, 2014; Hepper et al., 2011; Hughes & Beer, 2013; Korn et al., 2012). The sociometer hypothesis (Leary et al., 1995)

proposes self-esteem as a system for monitoring others' perceptions of us, providing an index of social valuation so that we can regulate our behaviors accordingly. Although these competing theoretical models conflict with one another at points, they are unified by the notion that self-esteem can be shaped by either selectively incorporating or ignoring the appraisals of others.

A body of research also suggests that there might be individual differences in the degree to which interpersonal perception influences self-esteem or vice versa. Although self-esteem is responsive to value cues from others (Weisbuch et al., 2009), these value cues seem to exert more influence on the self-view of individuals with lower self-esteem (Tice & Wallace, 2003), and those with less positive self-image engage in more perspective taking during self-reflection (Pfeifer et al., 2009) and are more likely to internalize peer rejection (Rodman et al., 2017). Focusing on the appraisals of others also leads to immediate drops in state self-esteem (Tice & Wallace, 2003). These findings suggest that incorporating the views of others into one's self-view can in fact be detrimental to self-esteem. In this view, having higher similarity between how we see ourselves and how our peers see us is associated with lower self-esteem.

Given these findings, it remains a possibility that distancing one's view of oneself from the views of others might enhance self-esteem, and indeed the literature on self-enhancement demonstrates this. Here, individuals maintain or boost their self-esteem by preferentially ignoring the negative appraisals of others (Anderson & Hanslmayr, 2014; Frey & Gaertner, 1986; Pyszczynski & Greenberg, 1985; Schröder-Abé et al., 2007), over-estimating positive appraisals (Hepper et al., 2011; Rodman et al., 2017; Somerville et al., 2010), or incorrectly interpreting ambiguous appraisals as self-flattering (Flagan et al., 2017). These findings suggest that possessing high self-esteem might be contingent on distancing one's self-worth from the

appraisals of others. Again, these findings support the idea that having lower agreement between how we see ourselves and how our peers see us would support higher self-esteem.

In recent years, there has been an increasing amount of research identifying the brain systems supporting self-evaluation and self-esteem. Though many regions are involved in social cognition more broadly, such as the cortical midline structures, temporal-parietal junction (TPJ) and the posterior cingulate cortex (PCC), a consistent finding across paradigms has identified the medial prefrontal cortex (MPFC) as a brain structure central to self-evaluation (Hughes & Beer, 2013; Somerville et al., 2010; Van de Groep et al., 2021; Will et al., 2017; Yankouskaya & Sui, 2021). The MPFC is a region that has been broadly linked to self-referential processing and person-perception processes (Wagner et al., 2012). These findings have since been extended to show that MPFC activation following social feedback can predict individual differences in self-esteem (Eisenberger et al., 2011; Somerville et al., 2010) and the MPFC seems to track evaluations from others (Ito et al., 2020) and prompt protective self-reflection in response to negative cues to buffer their debasing effect on self-esteem (Korn et al., 2012; Yoon et al., 2018). When the MPFC is disrupted with transcranial magnetic stimulation, there is a corresponding disruption in self-positivity bias, such disruption leads individuals to become more likely to attribute positive traits to a peer than themselves (Luber et al., 2012) and predict others will socially accept them at chance level rather than above chance (Zhang et al., 2022). Other research has linked individual differences in self-esteem to the connectivity of the MPFC with other brain systems linked to reward-related processes and positive affect such as the ventral striatum (Chavez et al., 2022; Chavez & Heatherton, 2015), suggesting that the MPFC is involved in integrating self-referential thought with evaluative processing. Taken together, this

body of evidence suggests that the MPFC is a critical brain region for understanding how the brain reflects self-esteem in social contexts.

Self/other reflection during interpersonal perception has been observed at the neural level. Using advances in multivariate functional magnetic resonance imaging (fMRI), previous work has found that the patterns of neural activity when an individual engages in self-reflection correlate with the patterns of neural activity of their peers reflecting on them (Chavez & Wagner, 2020). Put differently, this provides evidence that neural representations of the self are reflected, or recapitulated, in the brains of other in-group members. This effect—which we henceforth refer to as the *self-recapitulation effect*—was found in the MPFC identified from an independent functional localizer task. However, this initial study was conducted within a single social group of participants and did not have a sufficient sample size to detect individual differences. Thus, it remains an open question whether these findings extend to multiple and more diverse social groups and how individual differences in self-esteem may influence these results.

In the present study, we sought to inform theories of social influences on self-esteem by examining the relationships between self-esteem and self/other interpersonal perception processes at the level of the brain. To this end, we had two primary goals: first, we wanted to replicate the self-recapitulation effect using the round-robin fMRI methods from Chavez and Wagner (2020) but in a larger sample of participants drawn from multiple independent social groups and across larger brain networks implicated in social cognition. Second, we wanted to test whether individual differences in self-esteem were related to the strength of the self-recapitulation effect within the MPFC, given its established role in serving both social cognitive and self-evaluative processes. To this end, the current study used a multi-group round-robin design wherein previously acquainted individuals reflected on traits about themselves and their

group members while undergoing fMRI. Critically, the round-robin design required each individual to serve as both a target and perceiver for all other group members of their group. From here, we can compare an individual's neural activity patterns during self-reflection with the neural activity pattern of their peers reflecting on them. This comparison is the basis for calculating the self-recapitulation effect, or the agreement between brain-to-brain activity patterns between people thinking of themselves and group members thinking of them, which we can then relate to self-esteem.

Methods

Participants

We recruited a total of 120 right-handed participants from 20 independent real-world social groups, 6 people per group, from the University of Oregon and local Eugene community, including student organizations, local businesses, and friend groups. The sample size was determined by maximizing the number of groups and participants given the resources and feasibility of the design based on previous work (Chavez & Wagner, 2020) and is an order of magnitude greater than the previous study to increase power for detecting effects of self-esteem. All participants within each group were known to one another prior to entering the study. Across all groups, five participants failed to fulfill scheduled experimental session appointments, two subjects were excluded due to unusable imaging data, and three subjects were excluded for missing or incomplete behavioral data. After these exclusions, one group was eliminated for having too few subjects to calculate the recapitulation effect and the remaining three participants from that group were excluded. This left a final sample of 107 subjects across 19 groups in which all group members knew every other participant in the group. All participants were

screened for MRI contraindications and had normal or corrected to normal vision. Sample demographics for the participants are as follows: age (range = 18-51, $M = 23.50$, $SD = 7.36$), gender (Female = 39, Male = 68), race (American Indian or Alaskan Native = 2, Asian = 3, Black or African American = 5, Native Hawaiian or Pacific Islander = 2, White or Caucasian = 83, multiple = 10, other = 2), and ethnicity (Hispanic = 17, Non-Hispanic = 90). Participants gave informed consent in accordance with the guidelines set by the Internal Review Board at the University of Oregon and all were paid for their participation.

Procedures

Behavioral Measures

Participants completed two sessions: a behavioral-only session and an fMRI scanning session. In the first session, trait self-esteem was measured using the Janis-Field Feelings of Inadequacy scale (JFS, Fleming & Courtney, 1984), a thirty-six-item inventory of global self-esteem shown to have robust psychometric properties (Blascovich & Tomaka, 1991). The JFS asks participants to use a 5-point Likert scale to respond to a variety of items that assess self-evaluative feelings (e.g. “How often do you have the feeling that there is nothing you can do well?”, “When you think that some of the people you meet might have an unfavorable opinion of you, how concerned or worried do you feel about it?”, or “How often do you dislike yourself?”). Four items were reverse-scored and the scale was summed, with higher scores indicating higher feelings of inadequacy and lower self-esteem. Participants completed additional personality and mental health measures (Big Five Inventory, PROMIS mental health measures, stereotype content model) as part of a larger study on the brain basis of interpersonal perception (see:

Guthrie et al., 2022). All responses were recorded using the PsychoPy stimulus presentation software (Peirce, 2007).

Neuroimaging Procedures

Self-Referential Functional Localizer. In a second session, participants returned to complete the fMRI portion of the experiment, which, in replicating the protocol in Chavez & Wagner (2020), consisted of two standard versions of a trait-judgment task widely used in the study of self-referential processing (Mitchell et al., 2011). The first task participants completed in the scanner asked them to make trait evaluations for themselves or an “Ordinary American” as a functional localizer to independently define the region of MPFC that responds to self-evaluation to be used in subsequent analyses. Participants were presented with the word “SELF” or “ORDINARY AMERICAN” above a valence-balanced trait adjective and asked to respond “yes” or “no” using a button box as to whether the adjective described the individual above. The words were presented for 2,000 ms followed by 500 ms of fixation and intermittent passive fixation trials (2,500-7,500 ms). Jittered trials were optimized using Opseq2 (Dale, 1999). One run of this functional localizer task was completed and used to define a region of interest (ROI) within the MPFC that was maximally involved in self-referential processing and which was identified independently from the main task of interest. The functional localizer procedure is performed to avoid issues related to statistical non-independence or circular analysis which distort the interpretation of effects in neuroimaging studies (Kriegeskorte et al., 2009; Vul et al., 2009).

Main Self/Other Round-Robin Task. In this main task, we used a similar trait judgment task but modified to accommodate the round-robin nature of the interpersonal perception judgments we collected. Participants completed six runs of a round-robin version of a standard trait-judgment task, wherein participants were presented with a name from their small group alongside a trait-adjective and asked to respond via button-box as to whether this trait described that individual. For each trial, participants saw vertically-arranged white text on a black background with the top word either reading “SELF” or the name of one of the participant’s group members, and the bottom word displaying one of 360 valence-balanced trait-adjectives (e.g. “Happy”, “Clumsy”, “Smart”; Anderson, 1968) for 2000 ms followed by 2000 ms of fixation and jittered intermittent passive fixation trials (2000-12000 ms). Participants were asked to make either a “yes” or “no” response using a button box in their right hand as to whether or not the trait adjective described either themselves or one of their group members. All targets were presented in every run, and there were a total of twelve trials per target per run. Individual traits were only presented once per target randomly across all runs in the experiment. No two participants were presented with the same target/trait-adjective order across the experiment, to account for potential order effects.

Crucially, we employed a full round-robin design in which each participant was both a target and perceiver when asked to reflect on themselves along with all other members within their group. This design allows us to measure similarities in brain activity patterns to assess self-to-other similarity to test to what degree an individual’s neural activity pattern during self-reflection corresponds with the neural activity patterns of peers reflecting on the individual.

Image Acquisition

Magnetic resonance imaging was conducted with a Siemens 3T Skyra scanner using a 32-channel phased array coil. Structural images were acquired using a T1-weighted MP-RAGE protocol (175 sagittal slices; TR: 2500 ms; TE: 3.43 ms; flip angle: 7; 1 mm isotropic voxels). Functional images were acquired using a T2*-weighted echo planar sequence (TR: 2000 ms; 72 axial slices; TE: 25 ms; flip angle: 90; 2 mm isotropic voxels). For each participant, we collected one run of the functional localizer task and five runs of the main round-robin self-other task. In order to correct for distortion due to B0 inhomogeneity we also acquired a field map (TR: 6390 ms; TE: 47.8 ms; effective echo spacing: 0.345 ms). The total length of time for the entire scanning session was approximately 75 minutes.

Transparency and Openness

We report how we determined our sample size, all data exclusions and manipulations, and we follow JARS (Appelbaum et al., 2018). All data, analysis code, and research materials are available at <https://osf.io/n7pr2/>. This study's design and its analysis were not pre-registered.

Analyses

Functional Localizer Preprocessing and Analysis

All neuroimaging data including the functional localizer task were preprocessed and voxel responses were estimated using the FSL neuroimaging software package (Smith et al., 2004). For the functional localizer, data were first slice time corrected followed by a mean-based intensity normalization, high pass filtering (Gaussian-weighted least-squares straight line fitting, with $\sigma = 100$ s) and spatially smoothed with a 6-mm FWHM Gaussian smoothing kernel. First-

level analyses investigated the self > ordinary American contrast. These results were then submitted to a group-level random effects analysis to isolate regions of the brain involved in self-referential processing, correcting for multiple comparisons using voxelwise family error rate corrected p -values generated from $n = 5000$ iteration nonparametric permutation test. Note, the familywise error correction here refers to whole-brain voxelwise values and is more conservative than cluster-based correction or false discovery rate procedures. Voxels with a corrected p -value $< .05$ from this analysis were then converted into a binary mask that were then used at the group level to extract self and target response patterns within the round-robin task.

Round-Robin Task Preprocessing and Response Pattern Estimation

Preprocessing and registration of the main self–other task was identical to that of the functional localizer task described above except for the use of a smaller 4-mm FWHM Gaussian smoothing kernel (as is more typical of multivariate fMRI analyses where there is a need to preserve finer-scale neural responses). A multistep normalization procedure was used to register results to standard space. First, functional data were corrected for spatial distortions using a field map unwarping before aligning functional data to each participant’s anatomical scan using boundary-based registration (Greve & Fischl, 2009) in conjunction with a linear registration with FSL’s FLIRT tool. These images were then warped to a 2-mm MNI template using nonlinear registration with FSL’s FNIRT tool and a 10-mm warp field. All first-level task-based analyses were performed in the participant’s native space before being warped into standard space for final analyses. Parameter estimates were separately estimated for the self-condition as well as for each of the other peers within each of the six runs. These responses were then combined in a second level within-subject fixed-effects analysis yielding a parameter estimate for the self as

well as parameter estimates for each of the other group members participant's conditions. Normalized (i.e., z-score) voxel responses for each condition within the ROI defined in the functional localizer task served as input to the subsequent round-robin analysis.

Multivariate Round-Robin Task and Self-Esteem Analyses

Here, we aimed to measure the degree to which a target's activation pattern during self-trials correlates with aggregated activation patterns of group members reflecting on the target. For example, in a group with participants A, B, C, D, and E, a Spearman rank correlation of voxelwise parameter estimates within the regions identified by the functional localizer while A was thinking of A with the aggregate of voxelwise parameter estimates in the same ROI when B, C, D, and E were thinking of A. To this end, voxelwise parameter estimates within all areas identified by the functional localizer were flattened into one-dimensional response vectors to allow correlations across participants and conditions. Mirroring classic work in the interpersonal perception literature in which trait ratings are averaged across raters (McCrae & Costa, 1987), we used aggregated peer response patterns across voxelwise responses yielding a single group averaged response pattern for each target. As is standard in studies using fMRI to estimate multivariate similarity patterns (e.g., Kriegeskorte, 2008), we used Spearman rank correlation distance metric ($1 - \text{Spearman-rho}$) to compute the dissimilarity between neural response patterns during the self-condition and the group-averaged peer responses vector from the other participants in each group. Thus, this distance metric has the range of 0–2 with smaller values indicating greater similarity and higher agreement between self/other brain-to-brain activation patterns.

To provide evidence for the replication of the self-recapitulation effect, self-referential brain activity patterns within a target participant are correlated with those of others in their group when they are thinking of the target. However, rather than simply testing for a nonzero association between the target's self responses and their peers' responses, we sought to test the more conservative hypothesis that these relationships are specific to the given target and, in general, exceed that for other persons. To accomplish this, we also computed the correlation distance between a target's self-responses and the group response vector for every other participant within each social group. This allowed us to then calculate the similarity between self-congruent target comparisons (i.e., when perceivers were thinking of the same target as self) and self-incongruent target comparisons (i.e. when perceivers were thinking of a different target than the self). Importantly, the incongruent comparisons did not include the self-target responses so as not to influence the dissimilarity with respect to their own congruent patterns. Next, we fit a nested multilevel model comparing correlation distances between the self/other responses when congruent versus incongruent with random intercepts for each participant within each group.

Finally, we tested the hypothesis that the strength of the self-recapitulation effect (i.e. the self-congruent effects from the replication analysis) from each participant would be related to individual differences in self-esteem within the MPFC. To select a region of the MPFC for this analysis, we thresholded the functional localizer results to only include the most significant regions identified from that analysis $t(106) = 8.5$, such that the only voxels that remained in the mask were those within a contiguous cluster within the MPFC. The self-esteem correlation analyses were then calculated within this MPFC region for the self-esteem correlation analysis, as well as in the mask from the localizer as a whole. This was done in order to test the specificity

of the hypothesized results to the MPFC and not neural systems more broadly implicated in self-referential processing and social cognition.

Results

Replication of the Neural Self-Recapitulation Effect

Results from the functional localizer task identified a set of brain regions previously associated with self-referential processing and social cognition (see Figure 3.1A). These regions include the MPFC, precuneus/posterior cingulate cortex (PCC), temporal parietal junction (TPJ), lateral orbitofrontal cortex, and the ventral striatum. Within this ROI, we then calculated the correlation distance between a target's activation pattern during self trials and the aggregated activation patterns of group members reflecting on the target. This was iterated for every target in the round-robin design across all groups in the study to test for the self-recapitulation effect using a multilevel model to test whether the correlation distance between brain response patterns for the self was lower (i.e. more similar) when group members were thinking of the self (i.e. self-congruent correlation distance) than when group members were thinking of a different person in the group (i.e. self-incongruent correlation distance). The results show that other's self-congruent responses were more similar to the self-referential patterns than they were compared against others (see Figure 1B) after accounting for nested effects of groups and participants. These followed the same procedures as those from Chavez & Wagner (2020) but in multiple groups of subjects and within a significantly larger sample. These results replicate the main finding of evidence for the self-recapitulation effect in which brain response patterns for the self are reflected in aggregated responses from group members. Moreover, we find that these results

extend to show that these are present across multiple brain systems involved in both self-representation and social cognition.

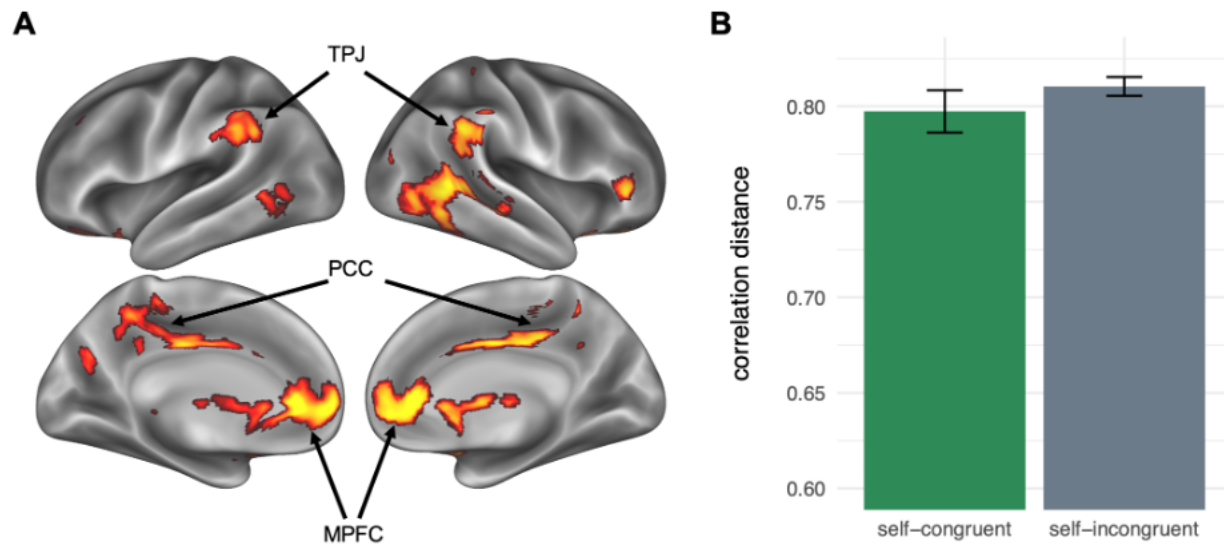


Figure 3.1. Results from the multi-group replication of the self-recapitulation effect from Chavez & Wagner (2020). (A) Regions identified from the functional localizer task in which brain activity when making trait-judgments of the self was greater than making trait-judgements for others ($p < .05$, voxelwise-corrected). Here we identified areas previously implicated in self-referential processing and social cognition including the medial prefrontal cortex (MPFC), posterior cingulate cortex (PCC), and the temporal-parietal junction (TPJ), and used these to independently define the brain areas examined in the round-robin task. (B) We show that self-referential brain activity for each target participant shows greater accuracy (lower correlation distance) with the brain activity of group members when they are thinking about that target (self-congruent) than when they are thinking of different targets (self-incongruent). For ease of interpretation, the results here are displayed as bar graphs depicting aggregated differences between conditions across the whole study, but inferences were made from data analyzed using a nested multilevel model accounting for group and participant effects.

Relationship of Self-Recapitulation Strength and Self-esteem

Next, we sought to determine whether the strength of the self-recapitulation effect from the multi-voxel fMRI patterns could be predicted by the target self-esteem. The most significant area identified in the functional localizer task identified a region of the MPFC that was used for the self-esteem correlation analysis (Figure 3.2A). Within this region, we found a significant

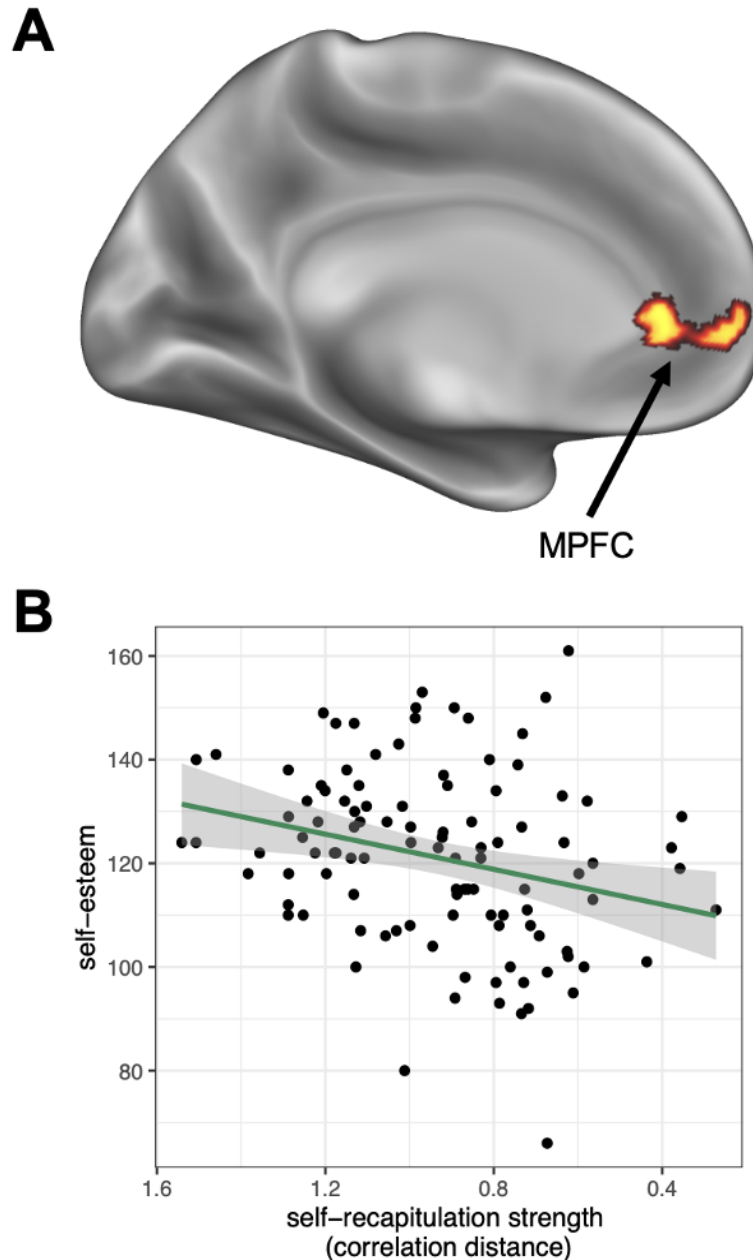


Figure 3.2. The results of the correlation of self-esteem with the strength of the recapitulation effect within the medial prefrontal cortex. (A) A portion of the medial prefrontal cortex (MPFC) that showed the most significant association with thinking of the self, relative to others from the functional localizer task and has been previously associated with self-evaluative processing. (B) A scatter plot depicting the correlation between self-recapitulation strength and self-esteem within the MPFC. These results show that as self-esteem increases self-recapitulation accuracy (i.e. correlation distance from the neural self-recapitulation analysis) decreases. The scale of the values on the x-axis of the plot is inverted to increase the clarity of the interpretation of correlation distance as an inverse measure of self-recapitulation strength from these analyses.

association between each individual's self-esteem and the similarity of patterns of self-representation to that of the aggregated group representations (i.e. the self-recapitulation strength), $R = .27$, $t(105) = 2.82$, $p = 0.006$ (Figure 2B). These results show that the lower a person's self-esteem, the more their neural representation within the MPFC self-reflection generally corresponds with those of their group members while reflecting on them.

In order to test the specificity of this effect to the MPFC, we repeated the same correlation procedure but within the functional localizer mask as a whole from the replication analysis. From this analysis, we find that there was no significant correlation between the self-recapitulation effect strength and self-esteem in broader regions of the social brain identified with the localizer task $R = .09$, $t(105) = 0.93$, $p = 0.35$. These results suggest that the effect of self-esteem modulates the self-recapitulation effect but only within the MPFC, which has previously been associated with self-evaluative processing, and provide increased specificity for the brain locations underlying this effect.

Discussion

We sought to test whether individual differences in self-esteem modulate the degree to which neural patterns overlap during self- and other-perception. To this end, we used a round-robin design wherein small groups of acquainted individuals reflected on traits about themselves and their group members while undergoing fMRI, and then determined the degree to which distributed patterns of brain activity in the MPFC during self- and other- reflection correlate. This design uniquely positions us to examine: 1) the degree to which an individual's neural activity pattern during self-reflection correlates with the neural activity patterns of peers

reflecting on the individual, and 2) whether individual differences in self-esteem predict the magnitude of this effect. Replicating the main results from Chavez and Wagner (2020), we find evidence for the self-recapitulation effect across multiple brain systems involved in both self-representation and social cognition. Furthermore, we also find that the self-recapitulation effect is moderated by self-esteem in the MPFC, such that lower self-esteem is associated with increased pattern overlap when that individual reflects on themselves and when their group reflects on them. These effects were specific to the MPFC, which has previously been identified as a region serving self-evaluative processes, and not found in other regions of the brain that serve other social cognitive processes. These findings have implications for both our understanding of the neural systems supporting interpersonal perception as well as social psychological theories of self-esteem.

A variety of competing theories have emphasized different primary sources of self-esteem. *Intrapersonal* perspectives on self-esteem propose self-esteem as a personal valuation of one's worth, in contrast to *interpersonal* perspectives that suggest that self-esteem is, at least in part, formed by the evaluation of others (Yang et al., 2016). Our findings suggest that there might in fact be individual differences in the degree to which an individual relies on an intra- or interpersonal model, such that self-esteem can be shaped by either selectively incorporating or ignoring the appraisals of others. To this end, we find that those who neurally represent themselves more similarly to how their peers represent them present with lower trait self-esteem. Although the present study was not designed to formally test theoretical models of self-esteem, our findings provide insight into these competing theories.

For example, some work has argued that higher self-esteem emerges from being perceived accurately by our peers (e.g. Self-Verification Theory; Swann & Read, 1981) with

other research suggesting the opposite – that higher self-esteem relies on us representing ourselves distinctly from how others view us, particularly in an inaccurately positive fashion (e.g. Self-Enhancement Theory; Krueger, 1998). Our findings provide a neural account for self-enhancement theories of self-esteem, such that individuals with higher self-esteem represent themselves more differently from how their peers represent them. This finding connects to the extensive literature on self-enhancement biases, which suggests that people tend to inflate their own self-worth even in the face of competing information (Anderson & Hanslmayr, 2014; Hepper et al., 2011; Hughes & Beer, 2013; Korn et al., 2012). Critically, we find this effect in the MPFC, and recent neuroimaging work has identified the MPFC as a brain structure that is core to the process of self-enhancement (Duran et al., 2020; Hughes & Beer, 2013; Van de Groep et al., 2021; Yankousaka & Sui, 2021), proposing the MPFC as a candidate region for the integration of social and self-evaluation that might compute self-protective responses in the face of negative social feedback (Will et al., 2017). The present study’s findings suggest that the neural integration of value cues from peers might render one more vulnerable to lower self-esteem and/or that individuals with lower self-esteem might more accurately integrate value cues from peers.

Limitations and Conclusions

Although we found evidence that lower self-esteem is associated with increased neural overlap between a target’s self-representation and how peers represent that target, these findings need to be interpreted in the context of their limitations. Firstly, our interpretations rest on the assumption that the trait-adjective experimental paradigm is reliably evoking self and other-concepts. While this task is regularly used to elicit interpersonal representations in neuroimaging

studies, the use of single trait adjectives might limit the richness of such representations. Even though this task represents one limited instantiation of self- and other- referential thought, the present findings suggest that similarities across even crude interpersonal representations are related to trait self-esteem. Moreover, while this task typically has individuals reflect on a distant other (e.g. a former president) or for varied targets across subjects (e.g. a parent), our round-robin paradigms allows all group members to reflect on the same, close others. We anticipate that conducting this task within real-world social groups increases the richness of person representations.

Our interpretations are also limited by the homogeneity of our sample, which was entirely adult and from the United States, and largely white. Conceptions of selfhood and the importance of self-esteem both vary across cultures (Goto, Lewis, & Grayzel-Ward, 2021; Gurel-Atay et al., 2010; Marcus & Kitayama, 1991; Triandis, 1989) and development (Robins & Trzesniewski, 2005), and our sample homogeneity limits our ability to generalize our findings to other populations. While cultural and developmental factors influence self-concept, across all humans, self-referential thought rests on psychological and biological processes at the level of the individual. We hope our findings can contribute to our understanding of such processes while also acknowledging how the homogeneity of our sample limits external validity. It is necessary that future work emphasize both cultural and developmental considerations into the examination of self-concept and self-esteem, given the context-laden nature of such constructs.

There are numerous theories of neural and cognitive processes underlying interpersonal perception, self-esteem, and the possible interaction of these two via the integration of others' perspectives into one's self-concept. Although our study was not designed to directly test competing self-esteem models, our results are most consistent with theories that state that

accurate self-perception is associated with lower self-esteem, such as sociometer theory or self-enhancement, rather than verification theories. Although future work is needed that explicitly tests out competing models of self-esteem using richer conceptions of self and other, the present findings suggest that similarities across even crude interpersonal representations are related to trait self-esteem. This reflects one possible neural mechanism underlying theoretical models that link self-esteem to individual differences in interpersonal perception.

CHAPTER IV

COMPUTATIONAL MECHANISMS OF SELF-ENHANCEMENT DURING SOCIAL COMPARISON AND THEIR RELATIONSHIP TO INTERNALIZING SYMPTOMS

This work was co-authored with R. Chavez and J. Clithero. I was the lead author of this publication and responsible for the writing of the manuscript with methodological details, and edits provided by co-authors. I led the study design, methods, and analyses, with input and assistance from co-authors.

Introduction

Cognitive biases in self-evaluation are a core, transdiagnostic feature of internalizing disorders, a descriptive label capturing diagnoses that feature persistent negative emotions directed inward, such as in unipolar mood and anxiety disorders (Eaton et al., 2013). Negative self-evaluation is listed in internalizing diagnostic criteria (DSM 5, APA, 2013; ICD, WHO, 2021), considered as an important part of the internalizing phenotype (Dixon et al., 2022; Ginot, 2012; Goldin et al., 2009; Nolen-Hoeksema et al., 2008; Talmon et al., 2021), and also central to prevailing theoretical models of internalizing disorders (Beck, 1967; Beck & Haigh, 2014; Clark & Wells, 2005). For example, Beck's (1967) seminal cognitive model posits that negatively biased self-knowledge is the main contributor to the onset and maintenance of depressive symptoms. The cognitive model remains essential to contemporary clinical research and practice, extending beyond depression to link negative self-evaluation with a variety of internalizing disorders (e.g. Beck & Haigh, 2014; Clark & Wells, 2005).

While positively biased self-evaluations contribute to well-being (Alessandri et al., 2014; Dufner et al., 2018; Schütz & Baumeister, 2017), research consistently demonstrates that negative self-evaluations are a feature of internalizing disorders. Various studies demonstrate elevated self-referential processing in depression (Kuiper & Derry, 1982; Lemogne et al., 2012; Li et al., 2017; Nejad et al., 2013; Northoff & Tumati, 2019), social anxiety disorder (Abraham et al., 2013; Lin et al., 2018; Yoon et al., 2019), and generalized anxiety disorder (Cui et al., 2022), particularly *negative* self-referential processing such as rumination and self-criticism (Castagna et al., 2019; Dinulescu et al., 2021; Ke et al., 2020; Kircanski et al., 2015; Nolen-Hoeksema et al., 2008). Internalizing symptoms are associated with enhanced memory for and processing of negative self-referential traits (Baños, Medina, & Pascual, 2001; Connolly, Abramson, & Alloy, 2015; Philippi, Cornejo, Frost, Walsh, Hoks, Birn, & Abercrombie, 2018). Importantly, while internalizing disorders feature a variety of symptoms, negative self-evaluations appear to be particularly influential for the severity, prognosis, and chronicity of these conditions (Disner, Schumake, & Beevers, 2017; LeMoult et al., 2016; Moscovitch et al., 2009; Nolen-Hoeksema et al., 2008; Orchard & Reynolds, 2018; Phillips, Hine, & Thorsteinsson, 2010), motivating the need to better understand such biases.

Previous work has studied the role of negative self-evaluation in internalizing disorders using information processing tasks such as the Self-Referent Encoding Task (SRET; Kuiper & Derry, 1982), which asks individuals to respond as to whether positive- or negative-trait words describe them or not. The SRET is believed to gauge the accessibility of positive or negative self-schemas, as measured by metrics such as response times, number of words endorsed as self-descriptive, and, and in some versions, word recall in a post-hoc memory task.

Research using the SRET has shown that individuals with depression, generalized anxiety, and social anxiety tend to make more negative self-endorsements and fewer positive self-endorsements (Davis, 1979; Dainer-Best et al., 2017; Dixon et al., 2022; Thurston et al., 2017; Tracy et al., 2021), and that non-depressed individuals consistently endorse more positive words than negative (Dainer-Best et al., 2017). Some work also shows that depressed individuals respond more quickly to negative self-endorsements (Alloy et al., 1997; Gilboa, Roberts, & Gotlib, 1997; MacDonald & Kuiper, 1985). However, other research finds no response time or endorsement rate differences (Collins & Winer, 2023; Kiang et al. 2017). These mixed findings likely reflect the fact that most research using the SRET relies on average response times and endorsements, coarse metrics that can obscure response dynamics. Moreover, such averages cannot identify the mechanisms driving altered responding – for instance, slower mean response times could reflect response bias, slower information processing, increased caution in decision-making, or slower perception or motor response (Pe et al., 2013). Given the relevance of aberrant self-evaluation to internalizing disorders, there is a need to evaluate SRET metrics with more granularity in order to resolve inconsistent findings and inform a mechanistic understanding of how they relate to clinical symptoms.

Recent calls to apply computational modeling techniques within clinical research (Poland & Frank, 2017) can be channeled to clarify the relationship between self-evaluation and internalizing symptoms. Computational approaches align with transdiagnostic research frameworks, better inform clinical interventions, and allow for more robust tests of psychology theories (Pitliya et al., 2022; Poland & Frank, 2017). One such approach is the Drift Diffusion Model (DDM; Ratcliff et al., 2016) which decomposes decisions into a set of latent components that reflect different aspects of the decision-making process. This makes DDM a promising

framework for studying self-referential biases, since they are often evaluated with binary forced-choice decision-making tasks like the SRET. While such tasks have demonstrated aberrant self-referential processing through mean response times, endorsements, or rate of recall, the DDM can help identify the cognitive processes driving such effects.

Emerging literature has applied computational modeling to the SRET (e.g. Beevers et al., 2019; Bränström et al., 2023; Castagna et al., 2023; Dainer-Best et al., 2018) under the assumption that decisions about whether traits are self-relevant involves evidence accumulation in support of one choice or another through the sampling of internal information from self-schemas (Hitchcock et al., 2023). This work has found that depression predicts a higher drift rate towards endorsing negative words as self-relevant (Beevers et al., 2019; Bränström et al., 2023; Dainer-Best et al., 2018; Castagna et al., 2023) and that psychological interventions can increase drift rate towards positive self-referents (Hitchcock et al., 2023). Despite the promise of this work, there are limitations. Most studies focus on depression symptoms, despite aberrant self-evaluation occurring across a range of internalizing disorders. Additionally, research is often constrained by pre-existing limitations of the SRET, like response asymmetry (Dainer-Best et al., 2018) and overlooking the social context in which self-evaluation occurs. More broadly, there is a need to replicate limited findings in this emerging area of inquiry. The present study aims to build on the small but growing literature applying DDM to self-referential decision-making tasks while addressing some of the aforementioned limitations.

In the present study, we use a modified version of the SRET, where, rather than responding *YES* or *NO* to indicate whether a trait is self-descriptive, participants instead respond as to whether the trait better describes themselves or their closest friend. This modification allows us to consider self-evaluation in the context of social comparison, which is important

because we have reason to believe that self-referential processing operates differently under such conditions (Kuiper & Rogers, 1979; Wagner et al., 2012). Early theories of self-evaluation highlighted the importance of social comparison (Festinger, 1954) with much research confirming that self- and other-evaluation are tangled processes (Chavez, 2020; Lieberman & Pfeifer, 2004; Wagner et al., 2012). Altered social comparison has also been identified as a potential mechanism underlying negative self-evaluations in internalizing disorders (McCarthy & Morina, 2020). While other papers have included both self and other as different conditions within the SRET (e.g. Kelley et al., 2002; Shestyuk & Deldin, 2010), these studies do not have individuals engage in explicit social comparison with peers within trials. An additional benefit of this modification is that it should help address the problem of response asymmetry in the SRET where individuals endorse YES for far more positive trials and fewer negative trials (Dainer-Best et al., 2018). It remains unclear whether such dramatic positivity biases persist in the context of social comparison with close peers.

Moreover, despite negative self-evaluation representing a transdiagnostic process related to multiple internalizing disorders, most work applying DDM to the SRET focuses on depression (e.g. Dainer-Best et al., 2018; Dainer-Best, 2023; Hitchcock et al. 2023). One recent paper examined DDM parameters in social anxiety disorder (Castagna et al., 2022), suggesting broader applicability across internalizing disorders. Further work is needed to determine how decision parameters may vary across different internalizing disorders. To this end, our paper examines the relationship between DDM parameters and multiple symptom profiles (depression, anxiety, social anxiety), adopting a dimensional approach to symptom severity. This method increases statistical power over approaches that categorize participants based on diagnostic criteria (Kotov et al., 2017) and builds on work that shows SRET metrics change across levels of

symptomatology (Timbremont & Braet, 2004; Timbremont, Braet, Bosmans, & Van Vlierberghe, 2008). It also could be possible that transdiagnostic measures might better capture the altered decision-making processes observed in internalizing disorders; with this aim, we also measured trait self-esteem. There is an emerging consensus around the need to consider self-esteem as a transdiagnostic process in order to better integrate our understanding of internalizing disorders, explain high rates of comorbidity, and better identify treatment treatments (Low et al., 2021).

In the present study, we use a modified SRET that situates self-evaluation in the context of social comparison, and examine how a variety of task metrics relate to both disorder-specific and transdiagnostic symptoms of internalizing disorders. We leverage developments from decision science, such as the HDDM Python toolbox (Wiecki et al., 2013), which uses hierarchical Bayesian estimation of DDM parameters. Our study builds on the small literature applying DDM to self-evaluative tasks by testing whether previous studies hold in the context of social comparison, across symptom profiles of multiple internalizing disorders, and when considering transdiagnostic factors such as self-esteem.

Methods

Sample

We recruited a total of 310 participants from the University of Oregon psychology undergraduate participant pool. Participants gave informed consent in accordance with the guidelines set by the Internal Review Board at the University of Oregon and all received course credit for their participation. Sample demographics for the participants are as follows: age (range

= 18-53, $M = 19.4$, $SD = 2.79$), gender (Female = 228, Male = 69, Non-Binary = 9, other = 2, decline to answer = 2), race/ethnicity (American Indian or Alaskan Native = 2, Asian = 32, Black or African American = 5, Hispanic = 36, Middle Eastern or North African = 2, Native Hawaiian or Pacific Islander = 2, White = 197, Multiple = 30, Other = 4). Participants were eligible to participate if they were over the age of 18, fluent in English, had normal or corrected vision, and provided informed consent.

Measures

Trait-Evaluation Task.

To measure self-evaluation, we used a trait-evaluation task modified from the SRET. The original SRET is a forced-choice decision-making task that presents participants with positive or negative trait-adjectives and asks them to judge whether the trait describes them or not in a binary fashion (yes/no; Derry & Kuiper, 1981). In order to examine whether self-referential processing biases still exist in the context of social comparisons, we deployed a modified version of the SRET where participants were instead asked as to whether each trait-adjective better described themselves or their closest friend (See figure 4.1). Trait words were either positively or negatively valenced, compiled from the Oregon Self-Concept Inventory (OSCI-2, Tucker et al., 2003) and other studies using trait-evaluation tasks (e.g. Grühn, 2016), and balanced for word frequency. Outcomes of interest include response times and endorsement rates, stratified by valence and response, as well as components of the DDM (discussed below).

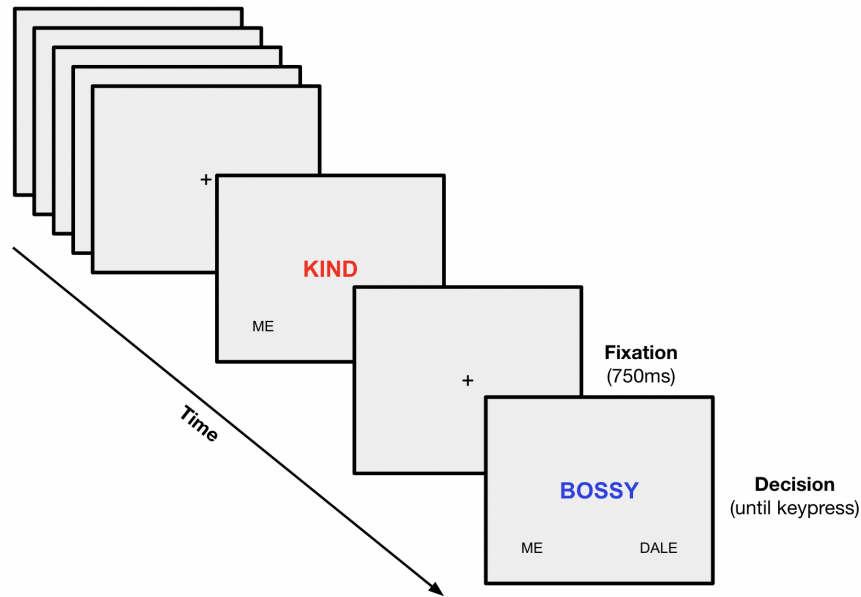


Figure 4.1. Schematic of trait-evaluation task. On each trial, participants were presented with a positively- or negatively-valenced trait-adjective and were instructed to respond via keypress (F or J) as to whether this word better describes themselves (“ME”) or a close friend, with the name of their friend presented on screen (here, “DALE”).

Janis-Field Feelings of Inadequacy Scale.

Trait self-esteem was measured using the Janis-Field Feelings of Inadequacy scale (JFS, Fleming & Courtney, 1984), a thirty-six-item inventory of global self-esteem shown to have robust psychometric properties (Blascovich & Tomaka, 1991). The JFS asks participants to use a 5-point Likert scale to respond to a variety of items that assess self-evaluative feelings (e.g. “How often do you have the feeling that there is nothing you can do well?”, “When you think that some of the people you meet might have an unfavorable opinion of you, how concerned or worried do you feel about it?”, or “How often do you dislike yourself?”). Four items were

reverse-scored and the scale was summed, with higher scores indicating higher feelings of inadequacy and lower self-esteem.

Patient-Reported Outcomes Measurement and Information System (PROMIS) - Emotional Distress - Depression.

Depression symptoms were measured using the PROMIS Item Bank v. 1.0 – Emotional Distress - Depression (Pilkonis et al., 2011), a set of 28 questions with strong psychometric properties (Cella et al., 2010) that is frequently used in clinical settings. The PROMIS - Depression asks participants to use a 5-point Likert scale indicating the frequency of their symptoms in the past 7 days (i.e. *Never, Rarely, Sometimes, Often, Always*). Questions probe a range of depressive symptoms, such as negative mood (e.g. “I felt sad”), anhedonia (e.g. “I felt that nothing was interesting”), cognitive deficits (e.g. “I had trouble making decisions”), negative views of the self (e.g. “I felt worthless”), and altered social cognition (e.g. “I withdrew from other people”; Pilkonis et al., 2011). Higher scores indicate higher levels of depressive symptomatology.

Patient-Reported Outcomes Measurement and Information System (PROMIS) - Emotional Distress - Anxiety.

Anxiety symptoms were also measured using the PROMIS Item Bank v. 1.0 – Emotional Distress - Anxiety (Pilkonis et al., 2011), a set of 29 questions that also possess strong psychometric properties and widespread clinical use (Cella et al., 2010). The PROMIS - Anxiety asks participants to use a 5-point Likert scale indicating the frequency of their symptoms in the past 7 days (i.e. *Never, Rarely, Sometimes, Often, Always*). Questions probe fear (e.g. “I felt frightened”), anxious misery (e.g. “I was concerned about my mental health”), hyperarousal (e.g. “I had difficulty calming down”), and physiological symptoms of arousal (e.g. “I had a racing or pounding heart”; Pilkonis et al., 2011).

DSM-5 Social Anxiety Disorder Severity Scale (SAD-D).

Social anxiety symptoms were measured using the DSM-5 Social Anxiety Disorder Severity Scale (SAD-D; LeBeau et al., 2012), a 10-item self-report measure examining the severity of social anxiety symptoms. The SAD-D is the only social anxiety scale based on the DSM-5 criteria (Wong et al., 2016) and was released by the APA in response to limitations of other measures by using items that directly correspond with diagnostic criteria and assess dimensionality rather than frequent dichotomous approaches to measuring social anxiety symptoms (i.e. social phobia vs. social interaction anxiety; Rice et al., 2021). Reliability and validity of the SAD-D have been established in both clinical and community samples (Binassis et al., 2021; LeBeau, Mesri, & Craske, 2016; Rice et al., 2021). Questions are rated on a five-point Likert scale (Never, Occasionally, Half the Time, Most of the Time, All of the Time) and probe cognitive & physiological symptoms related to social anxiety (e.g. “felt anxious, worried, or nervous about social situations,” “felt tense muscles, on edge or restless, or trouble relaxing”), socially-induced panic symptoms (e.g. “felt a racing heart, sweaty, trouble breathing, faint, or shaky in social situations”), and behavioral and cognitive avoidance (e.g. “distracted myself to avoid thinking about social situations”, “left social situations early or participated only minimally”).

Procedures

Upon arrival, participants were briefed about the purpose of the study and provided written informed consent. Each session was conducted in a quiet room with controlled lighting. First, participants completed computerized versions of the JFS, PROMIS-Depression, PROMIS-Anxiety, and SAD-D. Next, participants completed the trait-evaluation task on a 13 in x 8 in

monitor from approximately 23 in away. Before beginning the task, participants were provided with both verbal and written instructions for the task. They then completed 5 practice trials with the opportunity to ask questions. Next, participants completed a total of 264 trials with four brief breaks throughout the task. Each trial began with the presentation of a black fixation cross (750ms) on a gray screen. The fixation cross was replaced by a trait word at the center of the screen; below the trait word was the word ME on one side and the name of the participant's closest friend on the other. Whether each name was presented on the bottom left or bottom right was counterbalanced between participants to mitigate the confounding effect of handedness on response time. Participants were instructed to respond as quickly and accurately as possible as to whether the trait better described themselves or their close friend by clicking the F-key to indicate the name on the left side of the screen or the J-key to indicate the name on the right side of the screen, with a key-press terminating the trial. Trait words were presented once each in a randomized order. Stimuli presentation used PsychoPy (Peirce, 2007). After completing the task, participants were debriefed and given the opportunity to ask questions regarding the study.

Analysis

Data cleaning procedures were established according to previously established protocol (e.g. Dillon et al., 2015b). We removed outlier trials, defined as response times that were under 150 ms or above 3 standard deviations from the mean. If such outlier trials accounted for over 15% of a participant's total trials, we removed that participant from our sample. 15 participants were excluded because over 15% of their trials were outliers and 3 participants were excluded because they had no response variability (i.e. pressed the same button for every single trial). This resulted in a final sample of 292 participants for the following analyses.

Data Analytic Plan

Mean response time and endorsement rate analyses. In line with past work using trait-adjective tasks, we used response time (RT) and endorsement rates as dependent variables measuring how self-referential information is processed during decision-making (Dainer-Best et al., 2018). We operationalized RT as the average time from stimulus presentation until a keypress was made indicating a response. We calculated mean RTs for each participant across all trials, as well as each response type (i.e. mean RT when selecting *friend* vs. *self*), each trial type (i.e. mean RT for positive vs. negative trait-adjectives), and across each combination of condition and response (i.e. *self*/positive, *self*/negative, *friend*/positive, *friend*/negative). Endorsement rates were determined by summing the number of responses, providing a count of trials where participants endorsed *self* vs. *friend*, as well as a count of trials where participants endorsed *self* vs. *friend* for positive and negative trait-adjectives accordingly. We used these metrics as variables in linear models to predict mean-centered individual difference measures (i.e. self-esteem, depression, anxiety, and social anxiety) to determine if traditional trait-adjective metrics that measure self-schemas vary across internalizing symptomatology.

Mixed effects models. We conducted a series of linear mixed models using Restricted Maximum Likelihood (REML) to analyze response times and responses nested within condition with random intercepts for each participant. These models incorporated individual differences as random effects, and t-tests were employed with Satterthwaite's approximation for degrees of freedom to adjust for heterogeneity in variances. We ran models to test whether interaction effects between response and word valence predict RTs as well as additional models including an interaction term with each mean-centered individual difference measure (i.e. self-esteem, depression, anxiety, and social anxiety). We also ran mixed-effects logistic regressions to test

whether interactions between scaled individual difference measures and word valence predict responses. Here, we hoped to determine whether hierarchical models that preserve trial-level data can better predict RTs and endorsement rates than traditional metrics that collapse across trials.

Drift diffusion modeling. The DDM models binary forced choice decision-making tasks as a race from a starting point towards two decision bounds, where evidence in support of either decision accumulates until it reaches one of the bounds, at which point a response is made (Ratcliff & Koon, 2008). This model decomposes response times, responses, and their distribution into different components of the decision-making process. The typical DDM consists of four parameters: 1. starting point (z ; conceptualized as initial bias in favor of one response or the other and the starting point for evidence accumulation), 2. decision threshold (a ; conceptualized as the distance between decision boundaries and the amount of evidence required to make a decision, often indexing caution or motivation), 3. drift rate (v ; conceptualized as the rate of evidence accumulation, reflecting efficiency of evidence integration or the quality or strength of evidence supporting a particular choice), and 4. non-decision time (t ; conceptualized as non-decision factors, such as sensory perception or motor response).

We can consider the interpretation of each DDM parameter within the context of our study design (see Figure 4.2). In our forced-choice decision-making task, the lower decision bound reflected *friend* and the upper bound reflected *self*, with the distance between them being a . Accordingly, a larger a would mean that the decision thresholds were further apart, and thus a participant was being more deliberate or cautious when choosing between *self* and *friend*. The starting point closer to the upper bound would indicate that, independent of evidence accumulation, a participant has an initial bias towards responding *self*, a starting point closer to the lower point would indicate an initial bias towards responding *friend*, and a starting point in the middle between

each threshold would indicate no initial bias in responding *self* or *friend*. A positive drift rate would indicate more efficient evidence accumulation in the direction of the upper bound (*self*), and a negative drift rate indicates more efficient evidence accumulation in the direction of the lower bound, (*friend*), where drift rate is influenced by memory, value, and affective features of the decision. Non-decision time, while included in our model to account for sensory encoding and motor processes required of our task, was not a parameter of theoretical interest.

We ran a series of models using the HDDM Python toolbox (Wiecki et al. 2013), allowing a and v to vary according to word valence and individual difference measures with random intercepts included for each subject (see Table 4.3 for a list of models). For all models, 11,000 samples were generated, and the first 1,000 served as burn-in. Model convergence was determined by visually inspecting traces as well as calculating the Gelman-Rubin statistic (r-hat; Gelman & Rubin, 1992) which assesses the convergence of Markov Chain Monte Carlo (MCMC) simulations both within- and between- chains across different runs of the same model. Models with an r-hat ≥ 1.1 indicate model convergence, as an r-hat of 1.0 indicates that samples between chains are exactly the same. We used the Deviance Information Criterion (DIC) to compare different models, which calculates the deviance across the posterior distribution of the parameters, penalized for the number of model parameters. Thus, a lower DIC reflects a better model fit, accounting for both goodness of fit and model complexity.

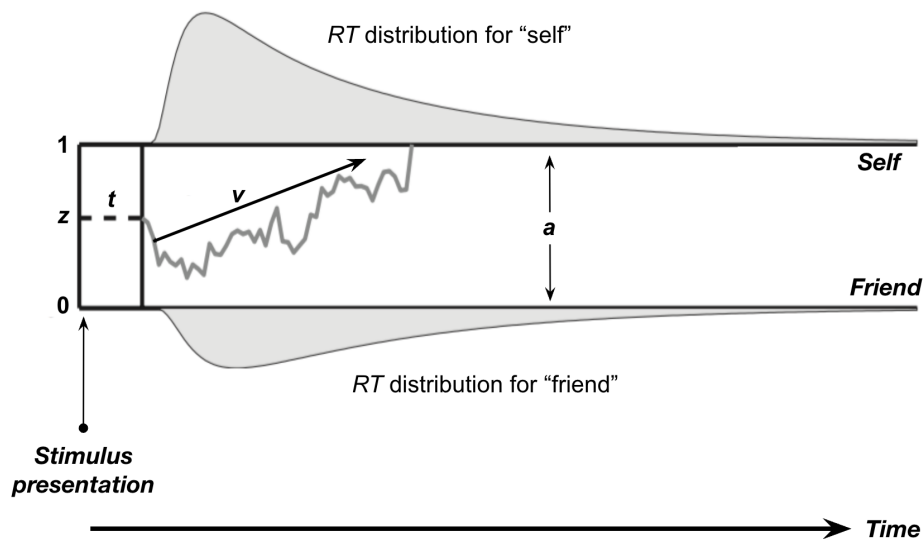


Figure 4.2. Schematic of the drift-diffusion model. In our task, an individual is presented with a positive or negative word (*stimulus presentation*) and then engages in a decision-making process to determine whether it better describes themselves (upper bound, 1) or a friend (lower bound, 0). The model decomposes the decision-making process into the following parameters: 1) z : initial starting point bias towards one decision bound or the other, 2) t : the time it takes to perceive and encode the stimulus (i.e. non-decision time), 3) v : the velocity of evidence accumulation towards a decision threshold, and 4) a : the distance between decision bounds, reflecting the quantity of evidence needed to enact a decision. Once enough evidence accumulates to reach a decision bound given the parameters of the model, a response is made. Figure adapted from Vandekerckhove, Tuerlinckx, & Lee (2010).

Results

Descriptive Statistics

RTs. RTs exhibited a mean of 1.50 s, a standard deviation of 0.85s, a median of 1.23s, were positively skewed (skewness = 1.82), and displayed excess kurtosis (kurtosis = 3.75). Mean response times across valence and response are reported in Table 4.1 and depicted in Figure 4.3.

Mental Health Measures. Despite recruiting from a college sample, a significant portion of participants reported elevated clinical scores. Of the final sample, 40.07% (N = 117) reported elevated depression scores according to the PROMIS Depression clinical cutoff (T-Score > 55; Cella et al., 2010), 17.12% in the mild range (N = 50), 20.55% in the moderate range (N=60), and 2.40% in the severe range (N = 7). 63.01% (N =184) had elevated anxiety scores according to the PROMIS Anxiety clinical cutoff (T-Score > 55; Cella et al., 2010), 14.73% in the mild range (N = 43), 39.73% in the moderate range (N=116), and 8.56% in the severe range (N = 25). 23.29% (N = 68) of the final sample had clinically significant levels of social anxiety symptoms according to the cutoff scores developed by Rice et al., (2021). All mental health measures were highly correlated with one another (Figure 4.6).

Mean Response Times

On average, participants responded more quickly when selecting *self* than *friend* ($t(291) = -8.48, p < 0.001$), and when responding to positively valenced words than negatively valenced words ($t(291) = -16.24, p < 0.001$). We aggregated trials by response and valence to examine differences across four combinations of valence and response: 1. negative/friend, 2. positive/friend, 3. negative/self, 4. positive /self. An ANOVA revealed a significant main effect

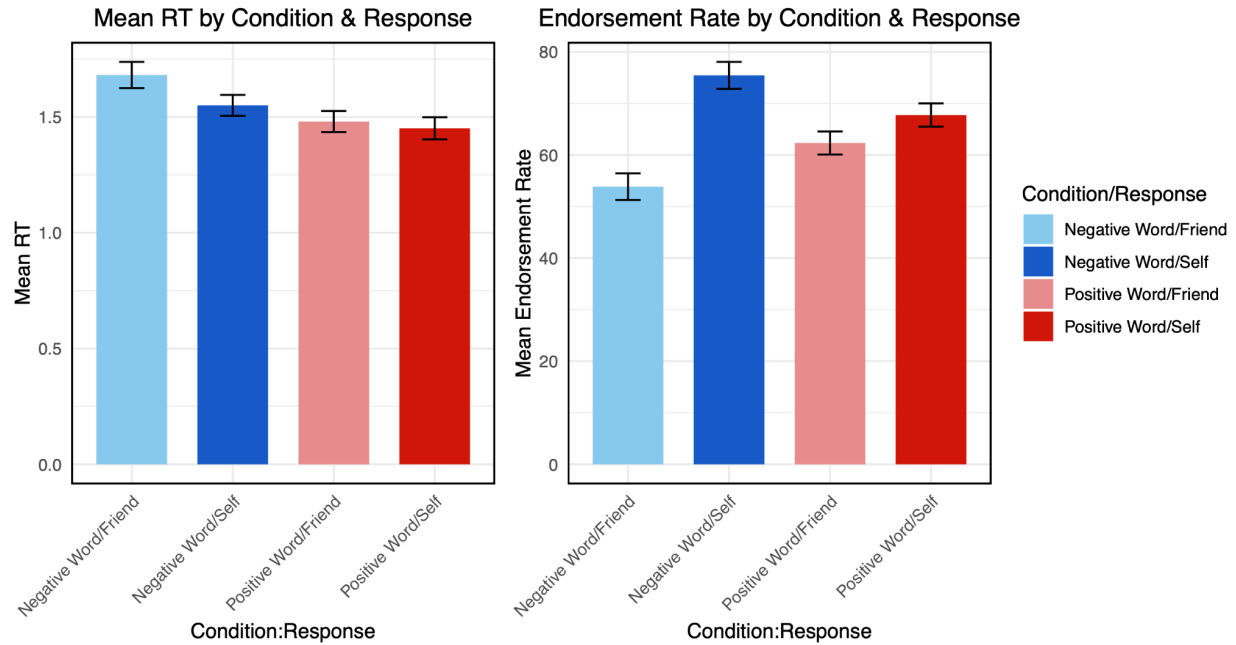


Figure 4.3. Mean RTs and endorsement rates for all participants averaged across valence and response. Mean RTs represents average response time in seconds. Mean endorsement rate represents a totaled count of choice (friend or self).

Table 4.1. Mean response times and endorsement rates by valence and response

Word Valence						
Response	Positive		Negative		Total	
	RT	Endorsement	RT	Endorsement	RT	Endorsement
Self	1.45 s	67.76	1.55 s	75.43	1.5 s	143.18
Friend	1.48 s	62.32	1.68 s	53.86	1.58 s	116.17
Total	1.47 s	130.07	1.62 s	129.29	1.54 s	—

of these condition/response combinations on mean response time ($F(3,873) = 88.11, p < 0.0001$, $\eta^2 = 0.042$). Mauchly's test indicated a violation of sphericity ($W = 0.847, p < 0.0001$), however significant effects remained after Greenhouse-Geisser ($\epsilon = 0.825, p < .0001$) and Huynh-Feldt corrections ($\epsilon = 0.832, p < .0001$). We used a linear mixed-effects model with post hoc contrasts of estimated marginal means and a Tukey correction for multiple comparisons. Participants' mean response times on negative/friend trials were significantly slower than on negative/self trials ($t(873) = 8.494, p < .0001$), positive/friend trials ($t(873) = 13.007, p < .0001$), and positive/self trials ($t(873) = 14.907, p < .0001$). Negative/self trials were slower than positive/friend ($t(873) = 4.513, p < .0001$) and positive/self trials ($t(873) = 6.414, p < .0001$). There was no significant difference in mean RT between positive/friend and positive/self trials ($t(873) = 1.9, p = 0.23$).

We ran linear regressions to examine whether mean-centered individual difference measures predicted differences in mean response times across each combination of condition/response. Worse self-esteem was associated with significantly slower mean RTs when selecting *friend* in response to a negative word ($b = 0.0042, SE = 0.0013, p < 0.001$) and *self* in response to a positive word ($b = 0.004, SE = 0.001, p < 0.001$). Higher social anxiety symptomatology was also associated with significantly slower mean RTs when selecting *friend* in response to a negative word ($b = 0.011, SE = 0.003758, p < 0.001$) and *self* in response to a positive word ($b = 0.009, SE = 0.03, p < 0.01$). Higher depression symptomatology was associated with faster mean RTs when selecting *self* on negative trials ($b = -0.002, SE = 0.0009, p = 0.035$). Anxiety symptoms did not significantly predict differences in mean response times across any combination of condition/response.

Endorsement Rates

On average, participants were more likely to respond *self* ($n=143.18$) than *friend*, $t(291) = 14.10$, $p < 0.011$. When looking at average endorsement rates across word valence and response, an ANOVA revealed that participants' average number of endorsements significantly differed across condition and response combinations ($F(3, 873) = 40.845$, $p < .0001$, $\eta^2 = 0.32$). Mauchly's Test for Sphericity shows a violation of sphericity for the condition ($W = 0.006$, $p < .0001$), however significant effects remain with both Greenhouse-Geisser ($GGe = 0.444$, $p < .0001$) and Huynh-Feldt ($HFe = 0.446$, $p < .0001$) corrections. Post-hoc contrasts with linear-mixed effects models using estimated marginal means and a Tukey correction for multiple comparisons revealed significant differences in endorsement rates between all response/valence combinations. On average, participants had fewer trials where they assigned a negative trait word to a friend than a negative trait word to themselves ($t(873) = -12.350$, $p < .0001$), a positive trait word to a friend ($t(873) = -4.844$, $p < .0001$) or a positive trait word to themselves ($t(873) = -7.951$, $p < .0001$). Additionally, on average, participants had significantly more trials where they selected *self* for a negative word than selecting *self* for a positive word ($t(873) = 4.399$, $p = .0001$) or friend for a negative word ($t(873) = 7.506$, $p < .0001$). Finally, when presented with a positive word, participants were more likely to endorse *self* than *friend* ($t(873) = -3.107$, $p = .0105$).

We ran linear regressions to examine whether mean-centered individual difference measures predicted differences in endorsement rates across each combination of condition/response. Worse self-esteem was associated with significantly more negative words endorsed for *self* ($b = 0.416$, $SE = 0.055$, $p < 0.0001$) and positive words endorsed for *friend* ($b = 0.4786$, $SE = 0.043$, $p < 0.0001$), as well as fewer positive words endorsed for *self* ($b = -0.496$,

SE = 0.043, $p < 0.0001$) and negative words endorsed for *friend* ($b = -0.4128$, SE = 0.0538, $p < 0.0001$). Higher depressive symptomatology was also associated with significantly more negative words endorsed for *self* ($b = 0.2319$, SE = 0.05293, $p < 0.0001$) and positive words endorsed for *friend* ($b = 0.22717$, SE = 0.045, $p < 0.0001$), as well as fewer positive words endorsed for *self* ($b = -0.234$, SE = 0.045, $p < 0.0001$) and negative words endorsed for *friend* ($b = -0.22531$, SE = 0.05232, $p < 0.0001$). Higher social anxiety symptomatology was also associated with significantly more negative words endorsed for *self* ($b = 0.9708$, SE = 0.1671, $p < 0.0001$) and positive words endorsed for *friend* ($b = 1.0683$, SE = 0.1376, $p < 0.0001$), as well as fewer positive words endorsed for *self* ($b = -1.092$, SE = 0.1384, $p < 0.0001$) and negative words endorsed for *friend* ($b = -0.9563$, SE = 0.1651, $p < 0.0001$). Finally, higher anxiety symptomatology followed this same pattern and was associated with significantly more negative words endorsed for *self* ($b = 0.22437$, SE = 0.05596, $p < 0.0001$) and positive words endorsed for *friend* ($b = 0.22191$, SE = 0.04749, $p < 0.0001$), as well as fewer positive words endorsed for *self* ($b = -0.22603$, SE = 0.04786, $p < 0.0001$) and negative words endorsed for *friend* ($b = -0.21809$, SE = 0.05531, $p < 0.0001$).

Mixed Effects Models

A linear mixed model predicting overall response times from word valence and response indicated that participants responded more quickly on positive trials ($b = -0.171$, SE = 0.009, $p < .0001$) and when responding *self* ($b = -0.101$, SE = 0.008, $p < 0.0001$). Additionally, there was a significant interaction between response and valence such that the faster response time observed on positive trials was less pronounced when a participant responded *self* rather than *friend* ($b = 0.068$, SE = 0.012, $p < 0.00001$; results summarized in Table 4.2).

Overall differences in response time were not predicted by any of our individual difference variables: self-esteem ($b = 0.0016$, $SE = 0.001$, $p = 0.114$), depression ($b = -0.00002$, $SE = 0.0009$, $p = 0.825$), anxiety ($b = -0.0002$, $SE = 0.00096$, $p = 0.802$), or social anxiety ($b = 0.0039$, $SE = 0.002937$, $p = 0.185$).

However, models including interaction terms revealed that interactions between mean-centered individual difference measures, word valence, and response type predicted response times (see Table 4.2 and Figure 4.4). In a model predicting response times from an interaction between mean-centered self-esteem, word valence, and response, we found that at mean levels of self-esteem, word valence had a significant negative effect on RTs ($b = -0.174$, $SE = 0.009$, $p < .0001$) as did the response (friend vs. self) ($b = -0.101$, $SE = 0.008$, $p < .0001$). Lower self-esteem increased RTs ($b = 0.0032$, $SE = 0.001$, $p = 0.0018$). Two-way interactions also predicted RTs: participants were slower to respond to *self* on positive trials than negative trials ($b = 0.073$, $SE = 0.012$, $p < .0001$), but as self-esteem worsened, participants were quicker to respond *friend* on positive trials ($b = -0.0025$, $SE = 0.0004$, $p < .0001$) and *self* on negative trials ($b = -0.00396$, $SE = 0.00037$, $p < .0001$). Finally, self-esteem moderated the relationship between valence and response such that the lower a participant's self-esteem, the slower they were to respond *self* on positive trials ($b = 0.00697$, $SE = 0.00054$, $p < .0001$).

Next, we ran a model predicting response times from an interaction between mean-centered social anxiety symptoms, word valence, and response. Results for social anxiety held the same pattern of significance and direction of effect as in the self-esteem interaction model and are summarized in Table 4.2.

We also ran a model predicting response times from an interaction between mean-centered depression symptoms, word valence, and response. Results mostly followed the same pattern as with self-esteem, except that increased depression symptoms did not significantly predict response time when selecting *friend* on negative ($b = 0.0076$, $SE = 0.0093$, $p = 0.42$) or positive trials ($b = 0.0064$, $SE = 0.00035$, $p = 0.0672$). Results followed this same pattern of significance and direction of effects when we ran a model predicting response times from an interaction between mean-centered anxiety symptoms, word valence, and response (results summarized in Table 4.2).

Hierarchical Bayesian Drift-Diffusion Modeling

We ran four drift diffusion models that estimated 1) all parameters at the group level (basic model), 2) v by valence condition, 3) a by valence condition, and 4) a and v by valence condition. Relative to a baseline model, adding valence condition improved model fit; the fourth model that included a regressor of valence on both drift rate and thresholds best fit our data ($DIC = 224669.85$; model fit summarized in Table 4.3). Consistent with findings from hierarchical models, participants had a starting point bias toward choosing *self* over *friend* ($z = 0.51$, $SD = 0.002$, $95\% CI = [0.50, 0.512]$, $MC_{err} = 0.00003$). Thresholds also varied by valence condition, with participants having smaller thresholds on positive trials ($a_{positive} = -0.17$, $SD = 0.01$, $95\% CI = [-0.18, -0.15]$, $MC_{err} = 0.0001$) reflecting more conservative decision-making when responding to negative over positive words. Valence impacted decision-making such that on positive trials, participants had a slower drift rate towards the *self* bound ($v_{negative} = 0.145$, $SD = 0.01$, $95\% CI = [0.127, 0.163]$, $MC_{err} = 0.00017$; $v_{positive} = -0.13$, $SD = 0.01$, $95\% CI = [-0.14, -$

0.12], $MCerr = 0.0001$). We used parameter estimates from the winning model for further Pearson's correlation analyses with individual difference measures.

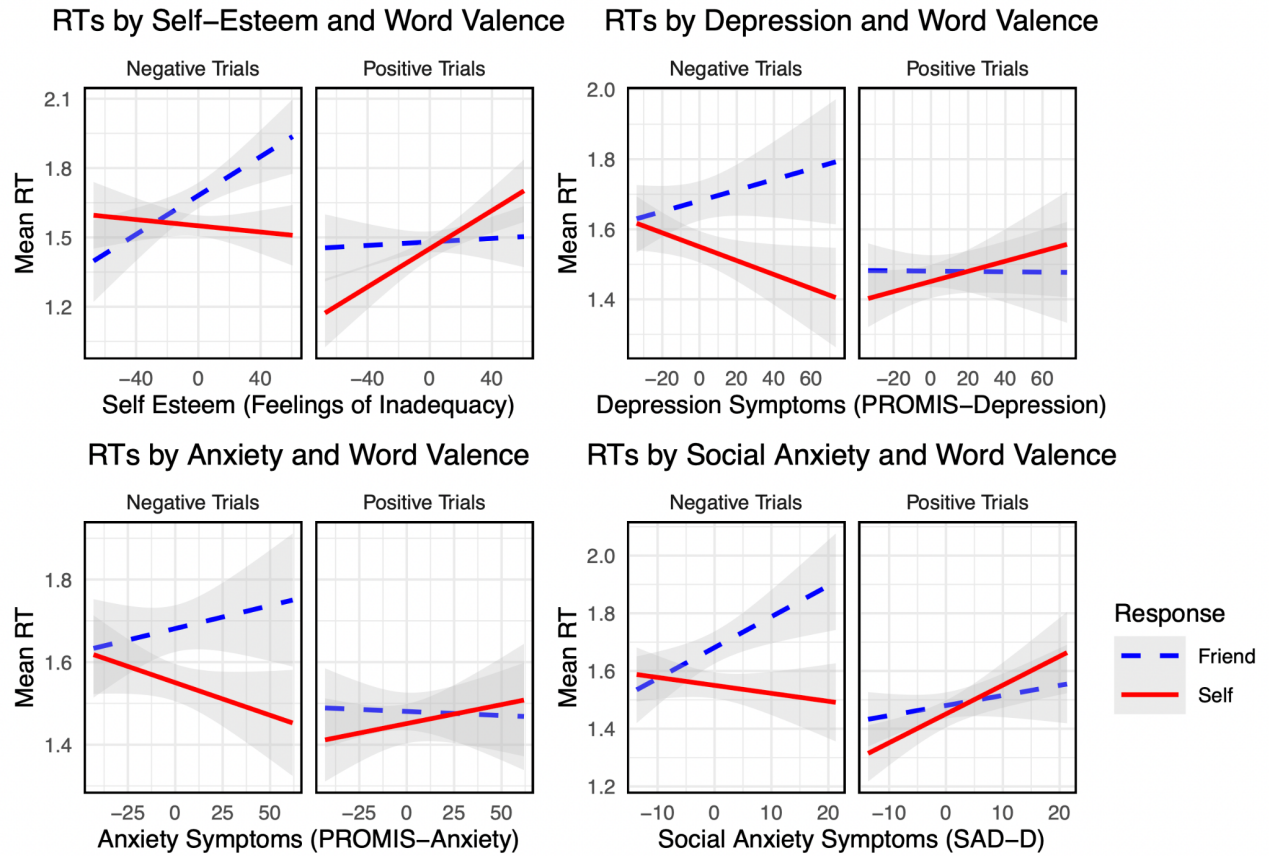


Figure 4.4. Mean response times plotted across word valence.

Table 4.2. Linear mixed effects models predicting response time

<i>Predictors</i>	RT			RT, <i>Scale</i>: Self-Esteem			RT, <i>Scale</i>: Depression			RT, <i>Scale</i>: Anxiety			RT, <i>Scale</i>: Social Anxiety		
	<i>b</i>	SE	<i>t</i>	<i>b</i>	SE	<i>t</i>	<i>b</i>	SE	<i>t</i>	<i>b</i>	SE	<i>t</i>	<i>b</i>	SE	<i>t</i>
(Intercept)	1.63***	0.02	71.29	1.63***	0.02	71.96	1.63***	0.02	71.37	1.63***	0.02	71.36	1.63***	0.02	71.76
Valence	-0.17***	0.01	-20.08	-0.17***	0.01	-20.31	-0.17***	0.01	-20.42	-0.17***	0.01	-20.29	-0.18***	0.01	-20.6
Response	-0.10***	0.01	-12.25	-0.1***	0.01	-12.19	-0.1***	0.01	-12.12	-0.1***	0.01	-12.15	-0.1***	0.01	-12.2
Valence x Response	0.07***	0.01	5.84	0.07***	0.01	6.25	0.07***	0.01	6.08	0.07***	0.01	5.96	0.07***	0.01	6.27
<i>Scale</i>	–	–	–	0**	0	3.16	0.001	0.001	0.81	0.001	0.001	0.55	0.01**	0	2.83
Valence x <i>Scale</i>	–	–	–	0***	0	-6.52	-0.001	0	-1.83	-0.001	0	-1.57	0.07***	0.01	6.27
Response x <i>Scale</i>	–	–	–	0***	0	-10.66	-0.003***	0	-8.16	-0.003***	0	-6.33	0.004***	0	-3.86
Valence x Response x <i>Scale</i>	–	–	–	0.01***	0	13.02	0.003***	0	7.73	0.003***	0	5.75	-0.01***	0	-10.03
AIC	174877.8			174682.4			174769.0			174823.7			174877.8		
BIC	174933.2			174744.8			174861.3			174916.1			174933.2		
Marginal R^2	0.008			0.012			0.009			0.011			0.008		
Conditional R^2	0.202			0.203			0.203			0.203			0.203		

Note. Above is a regression table describing the results of five mixed effects linear regression models, predicting response times (RT) from: 1) word valence, response and their interaction, 2) word valence, response, self-esteem as measured on the JFS, and their interactions, 3) word valence, response, depression symptoms as measured by the PROMIS-Depression and their interactions, 4) word valence, response, anxiety symptoms as measured by the PROMIS-Anxiety and their interactions, and 5) word valence, response, social anxiety symptoms as measured by the SAD-D, and their interactions). In all models, negative trials are coded as 0, positive trials as 1, friend responses are coded as 0, and self responses are coded as 1. All individual difference measures are mean-centered. $P < 0.001$ is indicated with ***, $p < 0.01$ is indicated with **.

Drift rate towards both positive and negative words significantly correlated with internalizing symptoms, such that participants who reported worse symptoms demonstrated a slower drift rate towards the *self* bound when responding on positive trials and a faster drift rate towards the *self* bound on negative trials (Figure 4.5). This held for all mental health measures: drift rate towards positive words was significantly correlated with self-esteem ($r = -0.48$, $p < 0.001$), depression ($r = -0.27$, $p < 0.001$), anxiety ($r = -0.25$, $p < 0.001$), and social anxiety ($r = -0.37$, $p < 0.001$); drift rate towards negative words was significantly correlated with self-esteem ($r = 0.40$, $p < 0.001$), depression ($r = 0.22$, $p < 0.001$), anxiety ($r = 0.20$, $p < 0.001$), and social anxiety ($r = 0.30$, $p < 0.001$; see Figure 4.5). Threshold and starting point bias did not significantly vary by individual difference measures. (Figure 4.6).

Table 4.3
Drift Diffusion Model Fits

Model	DIC	Deviance	pD
Basic model	225719.57	224736.21	983.35
v by valence	218510.871	217190.09	1320.78
a by valence	224766.55	223636.76	1129.79
a and v by valence	217426.67	215962.15	1464.52

Note. Model fits from four drift-diffusion models, estimating: 1) all parameters (a , t , v , z), 2) all parameters with v varying by word valence, 3) all parameters with a varying by word valence, 4) all parameters with both a and v varying by word valence. Model fit parameters include the Deviance Information Criterion (DIC), the posterior mean of deviance penalized for the number of model parameters that balances goodness of fit with model complexity. A lower DIC indicates better model fit. Deviance represents the amount of deviance in the specified model compared to a perfectly fitted model; thus lower deviance indicates better fit. Effective number of parameters (pD) captures model complexity with a higher value indicating model complexity. While the fourth model has the highest complexity, it best accounts for our data according to DIC and Deviance.

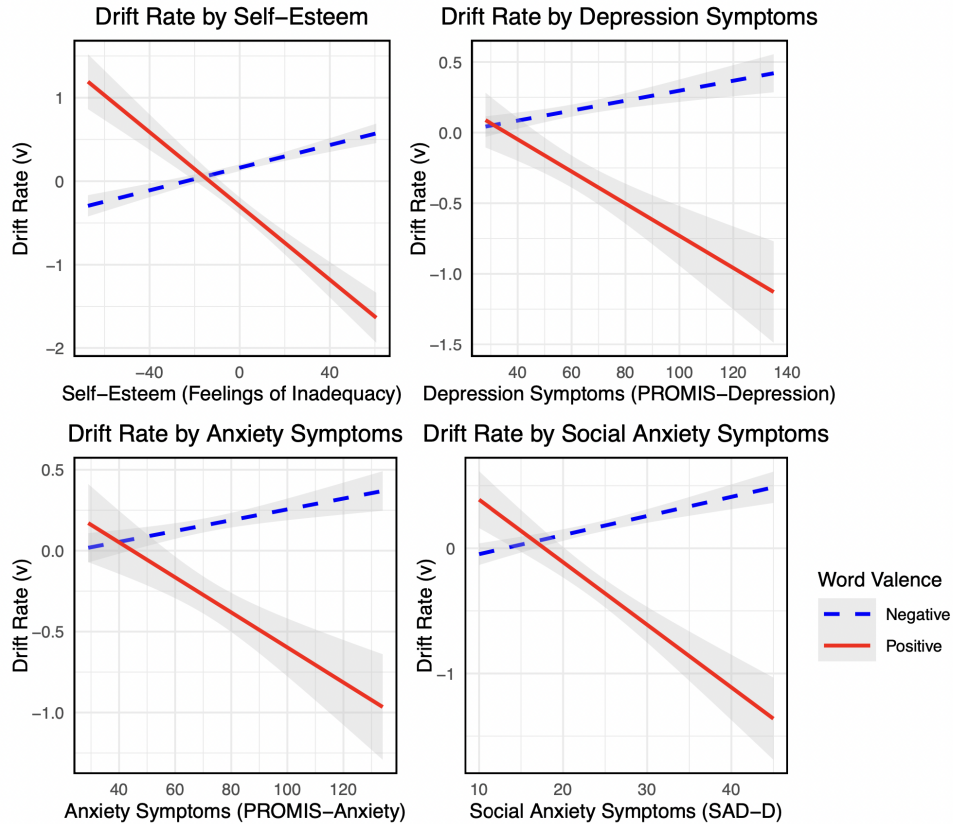


Figure 4.5. Drift Rate by Individual Difference Measures. Positive drift rates indicate increased efficiency of evidence accumulation towards the *self* decision bound and negative drift rates towards the *friend* decision bound.

Discussion

We applied drift diffusion modeling to a trait-adjective comparison task to determine if previous findings connecting computational parameters to internalizing symptoms are maintained in the context of social comparison. Our findings suggest that previously accepted ideas about self-enhancement may not apply when self-evaluation is situated within social comparison. For instance, while previous work using the SRET has shown that individuals are much more likely to endorse positive traits as self-relevant than not, here we find that social comparison disrupts this self-enhancement such that individuals are equally likely to attribute a positive trait to a friend as to themselves, and that the most frequently endorsed choice was

attributing *self* to a negative trait. Our more process-based analyses (RTs, DDM) show a mixed picture at the group level – while on average, individuals are slowest to respond *self* to negative traits, their drift rate is lowest on positive/self trials. Connecting these variables to individual differences is clarifying. We see that at higher levels of self-esteem, a small but persistent self-enhancement effect exists, where individuals more efficiently process evidence in favor of a positive self-concept. This effect flips as symptoms increase such that as individuals report more internalizing symptoms, they are more efficient in processing evidence in support of a negative self-concept, and vice versa for positive (Figure 4.5).

While both positive and negative drift rates significantly relate to all individual difference measures (Figure 4.5), larger effect sizes are consistently observed on positive trials. This finding is important because research, theory, and treatment of internalizing disorders disproportionality implicates negative self-evaluative biases (LeMoult & Gotlib, 2019). Our work confirms the importance of negative biases, showing that participants with internalizing symptoms more easily integrate information in support of negative self-schemas, it also demonstrates that difficulty retrieving positive self-schemas might be even more important. This fits in with parallel literature that observed dampened reward-processing in internalizing disorders, and further motivates the integration of clinical research on reward processes with work on biased self-referential processing. Treatments designed to increase the ease of retrieving positive self-schema could be an important avenue for further development (for emerging efforts in this area, see: Korrelboom et al., 2022; Bryant et al., 2023).

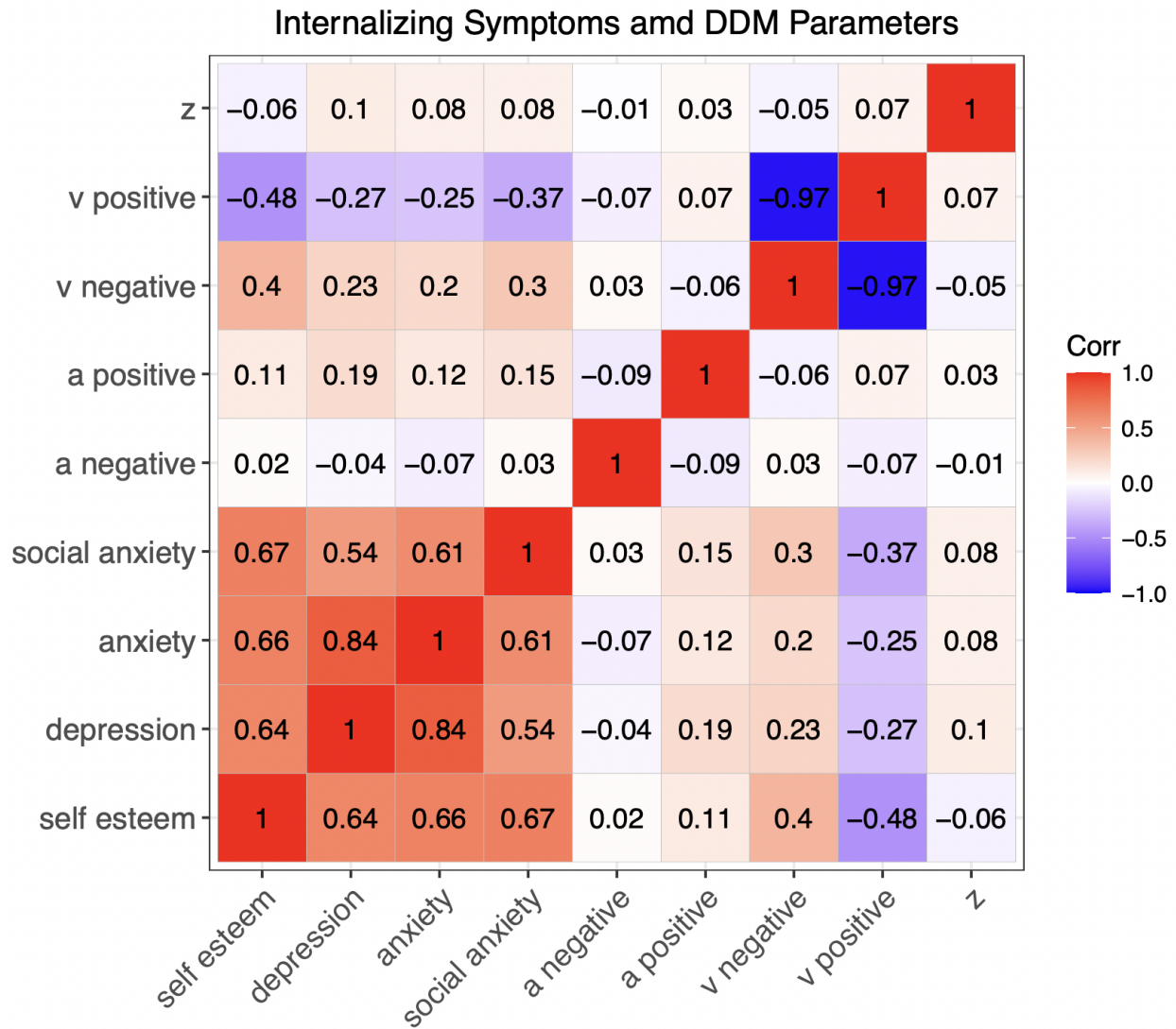


Figure 4.6. Correlation matrix of internalizing symptoms and DDM parameters (z , starting point bias; v , drift rate for positive and negative trials; a , decision bounds for positive and negative trials).

Our findings offer a mechanistic account of self-evaluative decision-making by showing that disruptions to evidence accumulation underlie the response biases observed in internalizing disorders. This finding is significant because it offers a more robust test of the psychological theories underlying biased self-evaluation. For example, depression might render individuals less motivated to choose self on positive trials, or anxiety might make individuals more cautious about responding correctly; these differences would manifest as differences in a across internalizing symptoms. The increases in self-referential processing in internalizing symptoms might bias individuals towards selecting self, which would show up as differences in z . Notably, we only find significant links between drift rate and internalizing symptoms, building on a growing body of work implicating evidence accumulation as the primary process driving aberrant, self-evaluative decision-making.

While examining DDM parameters affords us explanatory power, we found that drift rate had similar predictive power to endorsement rates, which replicates previous work (Disner et al., 2017). It is no surprise that endorsement rates would best predict self-reported internalizing symptoms. Binary choice trait-evaluation tasks have less measurement error than other metrics, and also might be collinear with mental health measures as both ask individuals to explicitly respond as to whether valenced descriptors are self-relevant. Given the overt nature of endorsements, this might also be where social desirability effects are most prominent (and thus, why we see attenuation of self-enhancement in endorsements during social comparison with well-liked peers). Computational modeling provides a measure of overt self-evaluation, where we observe that individuals with low levels of symptoms still show preferential processing of positive self-schema even when comparing oneself to a close peer. Given the barriers to computational modeling, if one is merely interested in evaluating explicit self-concept,

endorsement rates on the SRET might be sufficient. However, computational modeling offers insight into the mechanisms driving responses that are not evident in more overt measures, providing explanatory power and guiding treatment targets.

Most work applying DDM to self-evaluative tasks has focused on depression, even though biases in self-referential processing are observed across multiple internalizing disorders. We sought to build on work looking at self-evaluative processes in depression, and also include measures of social anxiety, anxiety, and the broader transdiagnostic construct of self-esteem. Our findings show that across internalizing diagnostic categories, we see relatively uniform patterns of disrupted self-evaluation. This suggests that changes to self-evaluation might reflect a transdiagnostic process occurring similarly across internalizing disorders. Supporting this is the finding that self-esteem is more strongly correlated with drift rate than the other mental health measures (Figure 4.6). We might find this because self-esteem might be a more sensitive measure to detect changes in performance on a self-evaluative task, rather than broader mental health measures which include items unrelated to self-evaluation. However, previous work has found symptom-level association between drift rate and the full spectrum of depression symptoms, including those with no apparent self-evaluative component (Hitchcock et al., 2023). This points to the utility of self-esteem as a transdiagnostic construct that might best capture changes to self-evaluation occurring across internalizing disorders. In line with increasing calls to study process-based models of clinical disorders (e.g. Kotov, 2017), future work studying abnormal self-evaluation could benefit from including transdiagnostic measures of self-esteem in addition to disorder-specific measures.

While we believe this work makes significant contributions to the self-evaluation literature, it also has limitations. Firstly, while this study aims to understand how self-evaluation

connects with internalizing symptoms, our use of a non-clinical sample limits our ability to generalize our findings to clinical populations. Even so, we have reason to believe our findings would replicate in clinical samples: college samples present with similar biases as clinical samples (Phillips, Hine, & Thorsteinsson, 2010), a high proportion of our sample reported scores above well-established clinical cutoffs, and dimensional models assert no qualitative differences between clinical and nonclinical populations (Kotov et al., 2017; Ruscio, 2019). There are also benefits to considering self-evaluation across a range of symptoms – here, we see changes emerge in a linear fashion as symptom levels increase, rather than drastically increasing after reaching a particular threshold. Still, future work is needed to replicate our findings in clinical samples confirmed by diagnostic interviews, and DDM’s strength in detecting between-group differences in model parameters makes it a compelling tool for comparing clinical groups to healthy controls. Future work can also block trials by word valence so that z can be modeled as a function of valence; while previous work using randomized trials still models z by running separate DDM for positive and negative trials (e.g. Dainer-Best et al., 2018; Castagna et al., 2023a, 2023b), we believe a blocked design would afford a more theoretically-grounded estimation of z and suggest future work block trials by valence. Future work can also make use of cutting-edge developments in sequential sampling modeling that allow for modeling regressors within the Bayesian framework, affording increased power and interpretability of other parameters. Finally, since this work showed that self-enhancement is altered in a social comparison task, future work can strive to better explain this effect by having individuals compare themselves to others across a range of closeness and/or liking.

In sum, we applied computational modeling to a modified valenced self-evaluation task that had individuals compare themselves to close peers. We observed attenuation of self-

enhancement effects in the context of social comparison, especially for more covert metrics revealed by the DDM. This attenuation in self-enhancement became more evident as internalizing symptoms increased, indicating that self-evaluative biases in internalizing disorders are best understood by reduced availability of positive self-schema, and to a lesser degree, increased availability of negative self-schema. This finding is an important contribution to a clinical literature that is predominantly focused on negative cognitive biases. It also bridges a large social psychology literature on self-enhancement with clinical work on self-referential processing.

CHAPTER V

GENERAL DISCUSSION

This dissertation wove together disparate literatures and methods to better understand how self-evaluation occurs within social contexts. Specifically, we drew on theory and methodology from social psychology, neuroscience, and computational psychiatry to identify possible mechanisms that link biased self-evaluation to internalizing disorders.

Chapter II integrated findings across social and clinical psychology to theorize about how socio-cognitive processes inform self-esteem. Various social psychology theories highlight the role of interpersonal perception in self-esteem and social relationships; yet often unacknowledged across these theories is the notion that *mentalizing* (i.e. inferring the mental states of others) shapes self-perception. This chapter integrated literature relating mentalizing to self-esteem, with a particular focus on *meta-perceptions*, the beliefs we hold about how others perceive us. We identified three core tendencies in meta-perceptions—accuracy, propensity, and permeability—that allow people to selectively integrate or ignore others’ appraisals in a manner that supports or diminishes self-esteem. We also proposed that different meta-perception tendencies can help explain low self-esteem as a risk factor for internalizing disorders, which are characterized by changes to self, social, and reward processes that can affect mentalizing to maintain symptoms.

In Chapter III, we explored the neural mechanisms that contribute to self-esteem within the rich social contexts that our review emphasized. Building on previous work demonstrating that neural representations of the self are reflected in the brains of peers (*Self-Recapitulation Effect*; Chavez & Wagner, 2020), we sought to explore how these processes are influenced by

individual differences in self-esteem. To this end, we used functional magnetic resonance imaging in a round-robin design to test how self-esteem modulates the representation of self-other similarity in multivariate brain response patterns during interpersonal perception. We replicated the self-recapitulation effect in a sample almost ten times the size of the original study and showed that these effects are found within distributed brain systems underlying self-representation and social cognition. Furthermore, we extended these findings to demonstrate that individual differences in self-esteem modulate these responses within the medial prefrontal cortex, a region implicated in both evaluative and self-referential processing. This work reconciles conflicting motives by supporting both the motive for self-knowledge (i.e. we strive to see ourselves as we are) and self-enhancement (i.e. we strive to see ourselves positively). Specifically, we find self-recapitulation effects across broad regions of the social brain that become increasingly dissimilar in the MPFC as self-esteem increases.

Chapter IV aimed to further unpack the cognitive mechanisms underlying self-enhancement biases in social contexts, with an eye to how these biases might shift across internalizing symptoms. Critically, we aimed to bridge clinical literature, which centers systematic negative self-evaluation as a core feature of internalizing disorders, with work from social psychology that highlights the critical role of positively biased self-evaluations in well-being. While previous studies have used the SRET to examine biased self-evaluation in internalizing disorders, none to date have examined self-evaluation in the context of social comparison while also using computational methods that provide mechanistic explanations for response differences. We modeled responses from a trait-evaluation task using drift-diffusion modeling (DDM) and related parameter estimates to individual differences in internalizing symptomatology. Our findings demonstrated that the relationship between internalizing

symptoms and altered self-evaluation is best explained by changes in the accumulation and integration of evidence in support of positive- and negative- self-views. This effect is particularly driven by responding on positive trials, where individuals with higher levels of symptomatology demonstrate more inefficiency in sampling representations that support a positive self-concept. These findings challenge assumptions about self-enhancement, demonstrating that 1) the self-enhancement effect is decreased in the context of social comparison, and 2) some individuals do *not* self-enhance, with this attenuation in self-enhancement being driven more strongly by disrupted evidence accumulation towards positive than negative self-schema.

Theoretical Implications

Self-Positivity

Together, our findings support the notion that healthy populations demonstrate self-positivity biases. We find that self-verification effects do not persist in brain regions that integrate information related to valence and self (Chapter III), and that in general, individuals are quicker to endorse positive-self information (Chapter IV). These results are aligned with well-established literature demonstrating self-enhancement (See Zell et al., 2020 for a recent meta-analysis and review). Moreover, our findings support the link between self-enhancement and positive affect & well-being (Alessandri et al., 2014; Dufner et al., 2018) in that we found higher levels of internalizing pathology predicted reduced positivity biases in self-evaluation.

Our findings confirm that self-enhancement confers some mental health benefits, reinforcing the adaptive and pervasive nature of self-positivity. Experiencing the self in a positive fashion is important for the will to survive and can serve as a motivating force towards

self-preservation. In this sense, self-positivity biases can be seen as the most atomic expression of in-group biases that have conferred evolutionary benefits over time. The closer an object, person, or idea is to our self, the more positively we tend to view it (Alves et al., 2016; Bialek et al., 2023; Shin et al. 2019) suggesting that self-other might represent a gradient of positivity. While the boundaries defining what is “ours” and “not ours” are consistently redrawn depending on the context, the self remains a locus of positivity. In this sense, self-enhancement biases represent what McKay & Dennet (2009) term an “inherent error”, a misbelief that represents an adaptive compromise, rather than a “doxastic dysfunction error”, a misbelief arising from dysfunctional cognition. As Karpen (2018) writes, “Our illusions occasionally mislead us, but they also get us out of bed,” conveying the adaptive nature of positively biased self-beliefs.

Empirical work confirms that self-enhancing biases offer adaptive benefits. Self-positivity improves our affect, increases our mastery over our physical environments, improves task performance, and motivates behavior toward self-identified goals (Alessandri et al., 2014; He & Côté, 2019; O’Mara & Gaertner, 2017; Sedikides, 2021; Wegner, 2003). Many studies demonstrate that self-positivity biases also protect us from psychological threats (Sedikides, 2021), echoing the “ego defenses” of classic psychoanalytic theories which were believed to reduce stress through denial (Freud, 1936). Our findings generally support the well-established effect that self-enhancement supports mental well-being (Alessandri et al., 2014; Dufner et al., 2018).

However, the work presented in this dissertation demonstrates that not everyone self-enhances, and that the lack of positive biases in self-view imposes higher costs than negative biases alone. In line with this, we found that while individuals with higher levels of internalizing symptomatology possess negatively biased self-evaluations, effect sizes were largest for

disturbed positive self-evaluations. This finding is important because most work examining self-referential cognition in internalizing disorders has primarily focused on *negative* self-evaluation (Beck, 1967; LeMoult & Gotlib, 2016).² While disruptions in positivity systems in these disorders have been well-established, they are typically explored through the lens of dysfunctional reward processing, motivation, and affect (e.g. Bogaert et al., 2023; Cataldo et al., 2023; Pool et al., 2016; Winer & Salem, 2016). Less work has directly examined how blunted positivity might also extend to self-evaluative processing. Our findings confirmed that as internalizing symptoms increase, negative-self information is processed more efficiently and positive-self information less efficiently, with the magnitude of effects being larger for the latter. Given this, along with the well-established link between positive self-evaluation and well-being in the social psychology literature, we strongly believe future work studying bias in internalizing disorders should consider *both* attenuated positivity and increased negativity. Moreover, studying attenuated self-positivity can bridge the social psychology literature on self-evaluation with clinical research that focuses on reward devaluation and blunted positivity.

Treatment Considerations

Treatments that address cognitive biases in internalizing disorders also tend to target negativity, likely due to the emphasis on negative self-evaluation in early models (Beck, 1967; Clark & Wells, 1995) and treatment protocols (e.g., Roiser et al., 2012). For example, Cognitive behavioral therapy (CBT) often targets negative self-evaluation, through cognitive restructuring or behavioral approaches that allow individuals to encounter information that challenge negative

² The “negativity bias” – the well-established finding that negativity is more salient than positive or neutral stimuli – might explain this asymmetric focus on negative self-evaluation in the literature. The one exception to the negativity bias is, unsurprisingly, the Self (Baumeister).

self-schema. While negative self-beliefs are frequently included as a study outcome, few studies measure positive self-beliefs (Gregory & Peters, 2017), despite good reason to believe CBT increases positive self-evaluation (Brewin, 2006; Kircanski et al., 2013). Other approaches include Cognitive Bias Modification (CBM), protocols that typically train individuals to avoid negative stimuli or interpret negative stimuli as neutral or positive (Kuckertz & Amir, 2017). Findings on the efficacy of CBM are mixed (Jones & Sharpe, 2017) and we wonder if targeting positivity in addition to negativity might lead to better results. Indeed, emerging work shows that explicitly targeting the processing of positive self-relevant stimuli improves outcomes (Blackwell et al., 2015; Craske et al., 2019; Dainer-Best et al. 2018; Korrelboom et al., 2022). According to our findings, future work should explicitly monitor and target both positive and negative biases to provide a more thorough account of their role in maintaining internalizing symptoms.

Our findings also provide mechanistic accounts of maladaptive self-evaluations, which can inform more precise and targeted interventions. Typical measurements of bias do not allow us to tease apart their underlying cognitive mechanisms, preventing us from identifying specific mechanisms as treatment targets. For example, many different cognitive processes could explain why an individual with depression might have a more difficult time making positive self-judgments; some possibilities include lower motivation to view oneself positively, impaired retrieval of memories in support of positive self-views, or a general cognitive bias towards negative stimuli over positive stimuli. The findings presented in Chapter IV suggest that disruptions in information integration processes best capture aberrant self-evaluation in internalizing disorders. Accordingly, treatments aimed at enhancing the processing efficiency of information that supports positive self-views might be more beneficial than general bias

modification. One such example is Competitive Memory Training (COMET; Korrelboom et al., 2022) which aims to strengthen the retrievability of representations (i.e. thoughts, memories) that support a positive self-view. Chapters II and III also suggest mentalizing processes might be an effective treatment target, for instance, encouraging individuals to distance their self-views from the opinions of others. Together, this work demonstrates that computational methods are valuable in their ability to inform clinical interventions and provide more robust tests of psychological theories.

Beyond the Self

While the above suggestions assume that increasing positive self-evaluation leads to uniformly beneficial outcomes, we cannot ignore that there are costs to self-enhancement, too. Even if self-enhancement makes us feel better, it can lead to worse relationships, poor school performance, unhealthy risk-taking, and increased prejudice (Baumeister, Heatherton & Tice, 1993; Colvin et al., 1995; Legault, Coleman, Jurchak, & Scaltas, 2021; Robbins & Beer, 2001). There is also evidence that self-enhancement is a less effective long-term strategy, leading to worse outcomes later on (Colvin et al., 1995; Leckelt, Küfner, Nestler, & Back, 2015), especially worse social outcomes (Dufner et al, 2018). There are also benefits to self-effacing. Humility has intrapersonal benefits (Exline & Hill, 2012; LaBouff, Rowatt, Johnson, Tsang, & Willerton, 2011; Van Tongeren, Davis, & Hook, 2014; Wright et al., 2017) and can support happiness (Zheng & Wu, 2020). Self-evaluations can also be shifted downwards in service of our goals (e.g. to strategically set expectations, Baumeister, Tice, & Hutton, 1989; to elicit social support, Leary, 2007). This more complicated picture might explain the failures of enthusiastic societal

efforts to increase self-esteem (e.g. 1986 California Task Force; Baumeister et al., 2005; Haney & Durlak, 1998; Mecca, Smelser, & Vasconcellos, 1989).

This begs the question as to whether it is beneficial to increase self-enhancement in individuals who do not already demonstrate this bias. An alternate path proposes reducing self-focus regardless of valence, given that elevated self-referential processing is a hallmark of internalizing psychopathology (Ingram, 1990; LeMoult, Kircanski, Prasad, & Gotlib, 2017). Research corroborates the benefit of reduced self-referential processing, showing reduced self-referential processing improves symptoms of depression (Lin, Callahan, & Moser et al., 2018; Tackman, Sbarra, Carey, et al., 2019). ‘Self-transcendent’ experiences (e.g. mindfulness, flow states, mystical-experiences) are associated with well-being (Yaden, Haidt, Hood, Vago, & Newberg, 2017), and are certainly integral to many spiritual, religious, and cultural technologies used for coping throughout history.

A Middle Path

Self-enhancement and -effacement both have costs and benefits, with the former generally improving interpersonal outcomes and the latter generally improving interpersonal outcomes (Humberg et al. 2018). Considering self-enhancement and self-effacement as strategies with relative tradeoffs opens the possibility that we can titrate them to serve particular goals. This view is aligned with broader research on resilience, where flexibility appears to support well-being more strongly than positive appraisals (Koole et al., 2015).

Flexibility means adjusting our self-evaluations based on context and goals. For example, a healthcare worker might find benefit in self-enhancing before giving a talk at grand rounds, but more accurately assessing their abilities while making life-or-death decisions for a

patient. And while the results of this dissertation demonstrated broad individual differences in the tendency to self-enhance, other work shows intra-individual flexibility is possible. For instance, one can minimize self-enhancement by reflecting on the strengths of others (Legault et al. 2021), perspective taking (Zhou et al., 2013), or evaluating oneself on specific, measurable, and modifiable criteria (Dunning, 1995). Further evidence of this flexibility is that self-enhancement and self-effacement can coexist: even when noting faults, individuals might overestimate their ability to improve those faults or the rate of those faults in the general population (Preuss & Alicke, 2017). On the flip side, even when individuals are rating themselves above average, they are also often comparing themselves to above-average targets (Davidai & Deri, 2019).³ Evidently, people already use both self-enhancement and self-effacement strategies, and perhaps we can learn to deploy them flexibly in service of our goals. Ultimately, self-evaluative tendencies that are too brittle can be detrimental.

Still, flexibility in self-evaluation strategies might be a lofty goal. Self-positivity biases are deeply ingrained (Karpen, 2018) and often beneath conscious processing (Chen et al., 2014; Chapter IV). For those with internalizing disorders, the grip of negative self-schemas can be powerful and hard to overcome. Perhaps the best we can do is mitigate the costs of such biases when they do emerge, rather than trying to correct them. Advances in self-compassion-based approaches show that changing our response to our biases – rather than the content of the biases themselves – can improve wellbeing (Ferrari et al., 2019). This approach encourages acceptance of our essential value, warts and all, in our messy human condition. Maybe the problem is not negative self-evaluations, especially those that help us see ourselves accurately, but rather allowing our unavoidable flaws to tarnish the intrinsic worth every human possesses.

³ Although, a warning about trying to self-enhance and self-efface at the same time: doing both at once (i.e. “humblebragging”) seems to evade the benefit of either (Sezer, Gino, & Norton, 2018).

Limitations and Future Directions

While our findings provide clarity into the relationship between self-evaluation, social cognition, and mental health, there are several constraints to the generalizability of the research and theories presented in this dissertation. A primary limitation is the underrepresentation of varied sociocultural perspectives. Conceptions of selfhood are culturally grounded (Goto et al. 2021; Markus & Kitayama, 1991; Triandis, 1989) and self-esteem is valued differently across cultures and geography, with research demonstrating that self-respect is especially important to individuals in the United States (Gurel-Atay et al., 2010). Although the present dissertation aimed to cite research from a range of disciplines, researchers, and countries, the literature we present is bound to a larger problem in the field of psychology, whereby the majority of research is conducted in and by W.E.I.R.D. populations (Western, Educated, Industrialized, Rich, and Democratic; Henrich et al., 2010). Moreover, our empirical chapters perpetuate this problem, as our samples were relatively homogeneous, being entirely adult and from the United States, and largely white. The authors that contributed to this dissertation are culturally bound themselves, coming from different research lineages and possessing both dominant and non-dominant identities across different dimensions of identity (e.g. gender, sexuality, race, ethnicity, religion, etc.). Indeed, our positionalities likely impact the nature of our research such that we are approaching culturally situated concepts, like selfhood, from a worldview that is common in Western, industrialized societies and thus will likely miss out on the nuanced experiences of selfhood as it arises across different cultures and subcultures. At the same time, despite clear and important cultural influences on the content and conceptualizations of selfhood, all humans must generate self-referential thought using psychological processes underpinned by biological mechanisms at the level of the individual. It is our hope that our work can speak to this level of

analysis while acknowledging the inferential boundaries and contextual limitations in which our conclusions are nested.

Given that most research on self-evaluation has been conducted in W.E.I.R.D. populations where confidence, autonomy, and power are culturally valued (Leary, 2007), the robust nature of the self-enhancement effect comes as no surprise. That being said, cross-cultural work shows that self-enhancement might be more universal than expected (Kim, Lee, Fo, Suh, Schimmack, et al., 2022; Kitayama & Uchida, 2003; Sedikides, Gaertner, & Toguchi, 2003, Sedikides, Gaertner, & Vevea, 2005). This includes studies conducted in cultures that strongly value humility, where self-enhancement effects still emerge, just less dramatically than in American samples (Shi et al., 2017) which might represent one extreme. Notably, this work has shown that while explicit measures of self-enhancement do not always hold cross-culturally, implicit measures still show self-enhancement biases (Shi et al., 2017); we observed this same effect in Chapter IV, where explicit measures of self-enhancement disappeared in the context of social comparison, but still were present in some of our implicit measures. Together, this suggests that self-enhancement might be a universal tendency that is differently expressed depending on cultural or social context. Others have also argued that enacting (or believing that one is enacting) traits that are valued in one's culture is still a form of self-enhancement, even if those traits are not the self-inflating traits valorized in U.S. culture (Leary, 2007). Regardless, the glorification of individualism in the United States serves as a major confound in self-evaluation research, and future work should strongly consider cultural considerations and sociological frameworks in the study of self and self-esteem. Future work might consider replicating existing findings cross-culturally, measure the personal and cultural value of evaluated traits, and

examine whether the costs and benefits of self-enhancement vary based on how a culture values self-esteem.

The presented empirical studies only sample adults, restricting our ability to examine developmental processes that might critically influence self-concept. Our work could benefit from a development lens since we draw connections between social, reward, and self-processing – domains that undergo fundamental shifts across development. “Self-conscious” emotions like shame, embarrassment, and guilt, only emerge when mentalizing capacities come online (Heery, Keltner, & Capps, 2003), reinforcing the proposed connection between social cognition, affect, and self-evaluation, as well as the importance of studying this link during relevant developmental periods. In particular, future work could examine whether our findings hold in the context of adolescence. Adolescence has been posited as a sensitive period for social cognitive development (Blakemore & Mills, 2014) when the social world becomes especially salient and self-concept increases in complexity (Shapka & Keating, 2005). This period is also characterized by increased reward sensitivity, especially to social rewards (Foulkes & Blakemore, 2016; Somerville & Casey, 2010). Notably, the incidence of internalizing disorders increases drastically during adolescence (Stattin & Skoog, 2023). I hypothesize that we see this increased incidence because the transition to adolescence introduces more sophisticated mentalizing and increased salience of social rewards that together can disrupt the positivity bias endemic to childhood (Thomaes, Brummelman, & Sedikides, 2017). This disruption might be adaptive in facilitating social learning and acquiring an accurate view of oneself required for increased independence, however can also increase the risk for internalizing disorders.

Another limitation of the present dissertation is the use of non-clinical samples. Chapters III and IV relied on a community sample and a convenience sample accordingly, where we

considered internalizing symptoms across a spectrum of severity based on self-report questionnaires. We did not collect data on formal diagnoses or conduct diagnostic interviews to establish whether participants met diagnostic criteria for disorders within the internalizing disorder spectrum or beyond. This limits our ability to generalize our findings to clinical populations and also prevents us from classifying those who scored low on our measures as ‘healthy controls’ as we did not evaluate other domains of psychopathology. At the same time, significant portions of our samples fell above clinical thresholds for depression, anxiety, and social anxiety disorder symptoms, providing some confidence that our beliefs would replicate in clinical samples. Further work is needed to replicate our findings using between-group paradigms that compare clinical groups with healthy controls. The DDM is well-suited to detecting between-group differences in response parameters and thus might be a particularly fruitful area for future investigation.

That being said, dimensional models of mental illness assert that there are no qualitative differences between clinical and nonclinical populations (Kotov et al., 2017; Ruscio, 2019). Dimensional approaches offer benefits, as they increase statistical power over approaches that categorize participants based on diagnostic criteria (Kotov et al., 2017) and allow us to detect changes across levels of symptomatology. Additionally, a meta-analysis shows that cognitive biases in depression do not differ between undergraduate samples from clinical samples (Phillips, et al. 2010), suggesting that some of the variables this dissertation is most concerned with might be appropriate for study within college samples. Moreover, in Chapter IV, all internalizing scales were highly correlated with one another (up to $R=0.84$), demonstrating broader problems with the validity of our current diagnostic classification symptoms which have long been subjects of debate (e.g. Poland et al., 1994). So, while it is pressing that future work

replicates our findings in clinical samples, our use of a dimensional model fulfills multiple aims of this dissertation, including studying transdiagnostic processes common to internalizing disorders and bridging research between social and clinical psychology.

The empirical chapters of this dissertation are also constrained by our study designs, which are both cross-sectional, lab-based, and use of trait-adjective tasks to study self-concept. Firstly, cross-sectional designs may obscure longitudinal relationships between self-evaluations and well-being. Given that previous work has found longer-term consequences of self-enhancement despite improved affect in the short term (Colvin et al., 1995; Dufner et al., 2018; Leckelt, Küfner, Nestler, & Back, 2015), it is critical that future work use longitudinal designs to clarify how the relationship between self-evaluation and affect evolves over time.

Secondly, our study designs are all lab-based. Additional research can benefit from increasing the ecological validity of paradigms, especially for research studying social processes where context is paramount. Although our studies took place in the lab, we used novel paradigms to study self-evaluation within real-life social relationships and our results support that self-evaluation operates differently within existing friendships and social comparisons. To this point, our findings challenge assumptions about self-enhancement effects using the SRET, where we do not replicate previous findings demonstrating asymmetric positive-self responses (Dainer-Best et al., 2018) under conditions of social comparison. However, our work could be complemented by creative paradigms with increased ecological validity; for example, future work could study self-referential biases on social media platforms, where self-evaluation is already occurring in social networks and influenced by peer comparison (e.g. Zheng et al., 2020).

Thirdly, our work is constrained by limits of trait-adjective tasks in reliably evoking self-concept. Though such tasks are standard in self-evaluation research, the use of single trait

adjectives might limit the richness of self-representations. Forthcoming work from our lab has used broader self-relevant narratives to evoke self-referential processing (Guthrie et al., forthcoming), which might capture more detailed representations. Another limitation of trait-adjective tasks is that they typically aggregate across broad categories (i.e. endorsed vs. not endorsed, positively vs. negatively valenced), obscuring individual differences in how those categories are perceived. For instance, some work shows that individuals are more likely to self-enhance when considering ambiguous traits (Dunning et al. 1989). Or, the personal or cultural value an individual places on a trait might matter more than its valence: do endorsements on valued traits contribute more to positive self-evaluation, even if those traits might be deemed ‘negative’? Might being perceived accurately on valued traits matter more for self-esteem? Considering the impact of value and reward is especially important, as this might reconcile conflicts in the literature between self-enhancement vs. self-knowledge motives. The results in Chapter III offer an account of both: there is neural overlap between representations of our self and our peer’s representations of us across broad regions of the social brain (self-knowledge), however, dissimilarity in the MPFC, which computes value, predicts self-esteem (self-enhancement). Future work that augments the trait-evaluation task to better account for reward and value-processing might be critical in clarifying conflicting findings.

One strength of our work is the focus on clarifying mechanisms that underlie aberrant self-evaluation. We identified that mentalizing might be a mechanism linking social information with self-evaluation (Chapter II), that neural dissimilarity between representations of our self and our peer’s representations of us in brain regions that integrate self- and value might underlie self-enhancement effects (Chapter III), and that disturbed evidence accumulation processes best account for disruptions in self-evaluation observed in internalizing disorders (Chapter IV). While

a plethora of work has shown the importance of self-esteem in internalizing disorders, we cannot actually target it without understanding its underlying mechanisms. Further research can build on existing theoretical and clinical work on internalizing disorders by applying methods from computational psychiatry to more robustly test theories, offer mechanistic insights, and better inform clinical interventions.

Conclusion

Thinking about ourselves in social contexts, either imaginally (Chapter II), alongside real-life peers (Chapter III), or in direct social comparison (Chapter IV), can shift our sense of self, often in ways that are detrimental to mental health. Across all chapters, we observe that situating ‘self’ within the ‘social’ disrupts self-positivity; internalizing the views of others (Chapters II & III) or engaging in social comparison (Chapter IV) both attenuate self-enhancement biases. These findings support a ‘tradeoff’ between the motive for self-enhancement and the motive for self-knowledge: the former supports us in feeling good within ourselves, while the latter supports us in relating well to others. Importantly, this dissertation argues that imbalance in these motives can be deleterious – both are effective strategies towards basic drives, and the ability to flexibly titrate between self-knowledge and self-enhancement might ultimately support the best outcomes. We believe that building such bridges across opposing theories can happen through the use of computational methods that offer more mechanistic accounts of the cognitive and neural processes underlying self-evaluation. We hope our findings make the case for studying self-evaluation with more nuance, whether that be by infusing classic theories with novel methodologies or testing them within the rich social context in which self-evaluation often occurs.

REFERENCES CITED

- Abraham, A., Kaufmann, C., Redlich, R., Hermann, A., Stark, R., Stevens, S., & Hermann, C. (2013). Self-referential and anxiety-relevant information processing in subclinical social anxiety: An fMRI study. *Brain Imaging and Behavior*, 7(1), 35–48. <https://doi.org/10.1007/s11682-012-9188-x>
- Ait Oumeziane, B., Jones, O., & Foti, D. (2019). Neural sensitivity to social and monetary reward in depression: Clarifying general and domain-specific deficits. *Frontiers in behavioral neuroscience*, 13, 199.
- Alessandri, G., Zuffianò, A., Fabes, R., Vecchione, M., & Martin, C. (2014). Linking positive affect and positive self-beliefs in daily life. *Journal of Happiness Studies*, 15, 1479-1493.
- Alicke, M. D., & Sedikides, C. (2011). *Handbook of Self-enhancement and Self-protection*. Allen, N. B., & Badcock, P. B. T. (2003). The Social Risk Hypothesis of Depressed Mood: Evolutionary, Psychosocial, and Neurobiological Perspectives. *Psychological Bulletin*, 129, 887–913. <https://doi.org/10.1037/0033-2909.129.6.887>
- Alloy, L. B., & Abramson, L. Y. (1988). Depressive realism: Four theoretical perspectives. In L. B. Alloy (Ed.), *Cognitive Processes in Depression* (pp. 223–265).
- Alloy, L. B., Abramson, L. Y., Murray, L. A., Whitehouse, W. G., & Hogan, M. E. (1997). Self-referent information-processing in individuals at high and low cognitive risk for depression. *Cognition & Emotion*, 11(5-6), 539-568.
- Alves, H., Koch, A., & Unkelbach, C. (2016). My friends are all alike—the relation between liking and perceived similarity in person perception. *Journal of Experimental Social Psychology*, 62, 103-117.
- American Psychiatric Association. (2015). *DSM-5®*. American Psychiatric Pub.
- Amir, N., Prouvost, C., & Kuckertz, J. M. (2012). Lack of a Benign Interpretation Bias in Social Anxiety Disorder. *Cognitive Behaviour Therapy*, 41(2), 119–129. <https://doi.org/10.1080/16506073.2012.662655>
- Anderson, M. C., & Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends in Cognitive Sciences*, 18(6), 279–292. <https://doi.org/10.1016/j.tics.2014.03.002>
- Andrews, P. W., & Thomson Jr., J. A. (2009). The bright side of being blue: Depression as an adaptation for analyzing complex problems. *Psychological Review*, 116, 620–654. <https://doi.org/10.1037/a0016242>
- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, 73(1), 3.

- Arch, J. J., Landy, L. N., Schneider, R. L., Koban, L., & Andrews-Hanna, J. R. (2018). Self-compassion induction enhances recovery from social stressors: Comparing adults with social anxiety disorder and healthy controls. *Anxiety, Stress, & Coping*, 31(5), 594–609. <https://doi.org/10.1080/10615806.2018.1504033>
- Back, M. D., Krause, S., Hirschmüller, S., Stopfer, J. M., Egloff, B., & Schmukle, S. C. (2009). Unraveling the three faces of self-esteem: A new information-processing sociometer perspective. *Journal of Research in Personality*, 43(5), 933–937. <https://doi.org/10.1016/j.jrp.2009.04.002>
- Ballespí, S., Vives, J., Sharp, C., Chanes, L., & Barrantes-Vidal, N. (2021). Self and Other Mentalizing Polarities and Dimensions of Mental Health: Association With Types of Symptoms, Functioning and Well-Being. *Frontiers in Psychology*, 12. <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.566254>
- Baños, R. M., Medina, P. M., & Pascual, J. (2001). Explicit and implicit memory biases in depression and panic disorder. *Behaviour Research and Therapy*, 39(1), 61-74.
- Batchelder, A. W., & Hagan, M. J. (2022). The Clinical Relevance of a Socioecological Conceptualization of Self-Worth. *Review of General Psychology*, 108926802211092. <https://doi.org/10.1177/10892680221109201>
- Baumeister, R. F., Tice, D. M., & Hutton, D. G. (1989). Self-presentational motivations and personality differences in self-esteem. *Journal of personality*, 57(3), 547-579.
- Baumeister, R. F., Heatherton, T. F., & Tice, D. M. (1993). When ego threats lead to self-regulation failure: negative consequences of high self-esteem. *Journal of personality and social psychology*, 64(1), 141.
- Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2005). Exploding the self-esteem myth. *Scientific American*, 292(1), 84-91.
- Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. New York: Harper & Row.
- Beck, A. T., & Weishaar, M. (1989). Cognitive Therapy. In *Comprehensive Handbook of Cognitive Therapy*.
- Beck, A. T., & Haigh, E. A. (2014). Advances in cognitive theory and therapy: The generic cognitive model. *Annual review of clinical psychology*, 10, 1-24.
- Beer, J. S. (2012). Self-evaluation and self-knowledge. *Handbook of Social Cognition*, 330-349.
- Beevers, C. G., Mullarkey, M. C., Dainer-Best, J., Stewart, R. A., Labrada, J., Allen, J. J., ... & Shumake, J. (2019). Association between negative cognitive bias and depression: A symptom-level approach. *Journal of abnormal psychology*, 128(3), 212.

- Berg, S. K., Bedwell, J. S., Dvorak, R. D., & Tone, E. B. (2021). Higher Social Anxiety Severity Predicts Better Cognitive Empathy Performance in Women but Not Men. *Psychological Reports*, 124(6), 2549–2566. <https://doi.org/10.1177/0033294120965496>
- Białek, M., Gao, Y., Yao, D., & Feldman, G. (2023). Owning leads to valuing: Meta-analysis of the mere ownership effect. *European Journal of Social Psychology*, 53(1), 90-107.
- Binas, T., Groves, D., Wootton, B. M., & Moses, K. (2022). Psychometric properties of the DSM-5 social anxiety disorder dimensional scale in an Australian community sample. *Journal of clinical psychology*, 78(5), 938-950.
- Blackwell, S. E., Browning, M., Mathews, A., Pictet, A., Welch, J., Davies, J., ... & Holmes, E. A. (2015). Positive imagery-based cognitive bias modification as a web-based treatment tool for depressed adults: a randomized controlled trial. *Clinical Psychological Science*, 3(1), 91-111.
- Blakemore, S. J., & Mills, K. L. (2014). Is adolescence a sensitive period for sociocultural processing? *Annual review of psychology*, 65, 187-207.
- Blascovich, J., & Tomaka, J. (1991). Measures of Self-Esteem. In *Measures of Personality and Social Psychological Attitudes* (pp. 115–160). Elsevier. <https://doi.org/10.1016/B978-0-12-590241-0.50008-3>
- Bleckmann, E., Lüdtke, O., Mueller, S., & Wagner, J. (2022). The role of interpersonal perceptions of social inclusion and personality in momentary self-esteem and self-esteem reactivity. *European Journal of Personality*, 08902070221080954. <https://doi.org/10.1177/08902070221080954>
- Bodenhausen, G. V., Kramer, G. P., & Süsser, K. (1994). Happiness and stereotypic thinking in social judgment. *Journal of Personality and Social Psychology*, 66, 621–632. <https://doi.org/10.1037/0022-3514.66.4.621>
- Bogaert, L., Dunn, B. D., Walentynowicz, M., & Raes, F. (2023). ‘It just doesn't feel right’ and other reasons why some people fear and avoid positive emotions. *Acta Psychologica*, 235, 103901.
- Bränström, R., Pachankis, J. E., Jin, J., Klein, D. N., & Hatzenbuehler, M. L. (2023). Self-schemas and information processing biases as mechanisms underlying sexual orientation disparities in depressive symptoms: Results from a longitudinal, population-based study. *Journal of psychopathology and clinical science*, 132(6), 681.
- Brausch, A. M., & Decker, K. M. (2014). Self-Esteem and Social Support as Moderators of Depression, Body Image, and Disordered Eating for Suicidal Ideation in Adolescents. *Journal of Abnormal Child Psychology*, 42(5), 779–789. <https://doi.org/10.1007/s10802-013-9822-0>

- Bress, J. N., Smith, E., Foti, D., Klein, D. N., & Hajcak, G. (2012). Neural response to reward and depressive symptoms in late childhood to early adolescence. *Biological psychology*, 89(1), 156-162.
- Bressi, C., Fronza, S., Minacapelli, E., Nocito, E. P., Dipasquale, E., Magri, L., Lionetti, F., & Barone, L. (2017). Short-Term Psychodynamic Psychotherapy with Mentalization-Based Techniques in Major Depressive Disorder patients: Relationship among alexithymia, reflective functioning, and outcome variables – A Pilot study. *Psychology and Psychotherapy: Theory, Research and Practice*, 90(3), 299–313. <https://doi.org/10.1111/papt.12110>
- Brewin, C. R. (2006). Understanding cognitive behaviour therapy: A retrieval competition account. *Behaviour research and therapy*, 44(6), 765-784.
- Briesemeister, B. B., Kuchinke, L., & Jacobs, A. M. (2012). Emotional valence: A bipolar continuum or two independent dimensions? *Sage Open*, 2(4), 2158244012466558.
- Brockner, J. (1988). *Self-esteem at work: Research, theory, and practice* (pp. xii, 258). Lexington Books/D. C. Heath and Com.
- Bryant, J., Winer, E. S., Salem, T., & Nadorff, M. R. (2017). Struggling toward reward: Recent experience of anhedonia interacts with motivation to predict reward pursuit in the face of a stressful manipulation. *PLoS ONE*, 12(3), e0173439. <https://doi.org/10.1371/journal.pone.0173439>
- Bryant, R., Dawson, K., Azevedo, S., Yadav, S., Tran, J., Choi-Christou, J., ... & Keyan, D. (2023). Positive affect training to reduce mental health problems during the COVID-19 pandemic: a proof-of-concept randomised clinical trial. *BMJ Ment Health*, 26(1).
- Burt, D. B., Zembar, M. J., & Niederehe, G. (1995). Depression and memory impairment: a meta-analysis of the association, its pattern, and specificity. *Psychological bulletin*, 117(2), 285.
- Butler, A. C., Hokanson, J. E., & Flynn, H. A. (1994). A comparison of self-esteem lability and low trait self-esteem as vulnerability factors for depression. *Journal of Personality and Social Psychology*, 6(1), 166. <https://doi.org/10.1037/0022-3514.66.1.166>
- Bylsma, L. M., Morris, B. H., & Rottenberg, J. (2008). A meta-analysis of emotional reactivity in major depressive disorder. *Clinical Psychology Review*, 28(4), 676–691. <https://doi.org/10.1016/j.cpr.2007.10.001>
- Byrne, G., Murphy, S., & Connon, G. (2020). Mentalization-based treatments with children and families: A systematic review of the literature. *Clinical Child Psychology and Psychiatry*, 25(4), 1022–1048. <https://doi.org/10.1177/1359104520920689>
- Byron, K., & Landis, B. (2020). Relational misperceptions in the workplace: New frontiers and challenges. *Organization Science*, 31, 223–242. <https://doi.org/10.1287/orsc.2019.1285>

- Cabooter, E., Millet, K., Weijters, B., & Pandelaere, M. (2016). The 'I' in extreme responding. *Journal of Consumer Psychology*, 26, 510–523.
<https://doi.org/10.1016/j.jcps.2016.03.002>
- Cataldo, A. M., Scheuer, L., Maksimovskiy, A. L., Germine, L. T., & Dillon, D. G. (2023). Abnormal evidence accumulation underlies the positive memory deficit in depression. *Journal of Experimental Psychology: General*, 152(1), 139.
- Campbell, J. D., Trapnell, P. D., Heine, S. J., Katz, I. M., Lavallee, L. F., & Lehman, D. R. (1996). Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *Journal of Personality and Social Psychology*, 70, 141–156.
<https://doi.org/10.1037/0022-3514.70.1.141>
- Cao, J., Gu, R., Bi, X., Zhu, X., & Wu, H. (2015). Unexpected Acceptance? Patients with Social Anxiety Disorder Manifest their Social Expectancy in ERPs During Social Feedback Processing. *Frontiers in Psychology*, 6.
<https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01745>
- Caouette, J. D., & Guyer, A. E. (2016). Cognitive distortions mediate depression and affective response to social acceptance and rejection. *Journal of Affective Disorders*, 190, 792–799. <https://doi.org/10.1016/j.jad.2015.11.015>
- Carlson, E. N. (2016). Do Psychologically Adjusted Individuals Know What Other People Really Think About Them? The Link Between Psychological Adjustment and Meta-Accuracy. *Social Psychological and Personality Science*, 7(7), 717–725.
<https://doi.org/10.1177/1948550616646424>
- Castagna, P. J., Calamia, M., & Davis, T. E. (2019). Negative self-statements mediate social anxiety and depressive symptomatology in youth: The role of working memory. *Journal of child and family studies*, 28, 1595-1603.
- Castagna, P. J., Waters, A. C., & Crowley, M. J. (2022). Computational modeling of self-referential processing reveals domain general associations with adolescent anxiety symptoms. *Research on Child and Adolescent Psychopathology*, 51(4), 455-468.
- Castagna, P. J., Waters, A. C., Edgar, E. V., Budagzad-Jacobson, R., & Crowley, M. J. (2023). Catch the drift: Depressive symptoms track neural response during more efficient decision-making for negative self-referents. *Journal of affective disorders reports*, 13, 100593.
- Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., ... & PROMIS Cooperative Group. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of clinical epidemiology*, 63(11), 1179-1194.
- Chavez, R. S., & Heatherton, T. F. (2015). Multimodal frontostriatal connectivity underlies individual differences in self-esteem. *Social Cognitive and Affective Neuroscience*, 10(3), 364-370.

- 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.
- Chavez, R. S., & Wagner, D. D. (2020). The neural representation of self is recapitulated in the brains of friends: A round-robin fMRI study. *Journal of Personality and Social Psychology*, 118(3), 407–416. <https://doi.org/10.1037/pspa0000178>
- Chavez, R. S., Tovar, D. T., Stendel, M. S., & Guthrie, T. D. (2022). Generalizing effects of frontostriatal structural connectivity on self-esteem using predictive modeling. *Cortex*, 146, 66–73. <https://doi.org/10.1016/j.cortex.2021.09.021>
- Chen, Y., Zhong, Y., Zhou, H., Zhang, S., Tan, Q., & Fan, W. (2014). Evidence for implicit self-positivity bias: an event-related brain potential study. *Experimental brain research*, 232, 985-994.
- Chepenik, L. G., Cornew, L. A., & Farah, M. J. (2007). The influence of sad mood on cognition. *Emotion (Washington, D.C.)*, 7(4), 802–811. <https://doi.org/10.1037/1528-3542.7.4.802>
- Clark, D. M., Crozier, W. R., & Alden, L. E. (2005). A cognitive perspective on social phobia. *The essential handbook of social anxiety for clinicians*, 193-218.
- Collins, A. C., & Winer, E. S. (2022). Self-referential processing and depression: A systematic review and meta-analysis. *Clinical Psychological Science*, 21677026231190390.
- Collins, A. C., Lass, A. N., & Winer, E. S. (2023). Negative self-schemas and devaluation of positivity in depressed individuals: A moderated network analysis. *Current Psychology*, 42(36), 32566-32575.
- Colmsee, I.-S. O., Hank, P., & Bošnjak, M. (2021). Low Self-Esteem as a Risk Factor for Eating Disorders. *Zeitschrift Für Psychologie*, 229(1), 48–69. <https://doi.org/10.1027/2151-2604/a000433>
- Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: negative implications for mental health. *Journal of personality and social psychology*, 68(6), 1152.
- Connolly, S. L., Abramson, L. Y., & Alloy, L. B. (2016). Information processing biases concurrently and prospectively predict depressive symptoms in adolescents: Evidence from a self-referent encoding task. *Cognition and Emotion*, 30(3), 550-560.
- Converse, B. A., Lin, S., Keysar, B., & Epley, N. (2008). In the mood to get over yourself: Mood affects theory-of-mind use. *Emotion*, 8, 725–730. <https://doi.org/10.1037/a0013283>
- Cooley, C. H. (1902). *Human Nature and the Social Order*. Charles Scribner's Sons. <https://books.google.com/books?hl=en&lr=&id=abo7EABChYAC&pgis=1>

- Craske, M. G., Meuret, A. E., Ritz, T., Treanor, M., Dour, H., & Rosenfield, D. (2019). Positive affect treatment for depression and anxiety: A randomized clinical trial for a core feature of anhedonia. *Journal of consulting and clinical psychology*, 87(5), 457.
- Cui, Q., Chen, Y., Tang, Q., Sheng, W., Li, D., Zeng, Y., Jiang, K., He, Z., & Chen, H. (2022). Neural mechanisms of aberrant self-referential processing in patients with generalized anxiety disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 119, 110595. <https://doi.org/10.1016/j.pnpbp.2022.110595>
- Cusi, A. M., Nazarov, A., MacQueen, G. M., & McKinnon, M. C. (2013). Theory of mind deficits in patients with mild symptoms of major depressive disorder. *Psychiatry Research*, 210(2), 672–674. <https://doi.org/10.1016/j.psychres.2013.06.018>
- Dainer-Best, J., Trujillo, L. T., Schnyer, D. M., & Beevers, C. G. (2017). Sustained engagement of attention is associated with increased negative self-referent processing in major depressive disorder. *Biological Psychology*, 129, 231–241. <https://doi.org/10.1016/j.biopsycho.2017.09.005>
- Dainer-Best, J., Lee, H. Y., Shumake, J. D., Yeager, D. S., & Beevers, C. G. (2018). Determining optimal parameters of the self-referent encoding task: A large-scale examination of self-referent cognition and depression. *Psychological Assessment*, 30(11), 1527.
- Dainer-Best, J., Shumake, J. D., & Beevers, C. G. (2018). Positive imagery training increases positive self-referent cognition in depression. *Behaviour research and therapy*, 111, 72–83.
- Dainer-Best, J. (2023). Trends in Research on Self-Reference, Depression, and the Self-Referent Encoding Task: A Critical Review. Preprint.
- Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping*, 8(2–3), 109–114. [https://doi.org/10.1002/\(SICI\)1097-0193\(1999\)8:2/3<109::AID-HBM7>3.0.CO;2-W](https://doi.org/10.1002/(SICI)1097-0193(1999)8:2/3<109::AID-HBM7>3.0.CO;2-W)
- D'Argembeau, A., Jedidi, H., Baeteau, E., Bahri, M., Phillips, C., & Salmon, E. (2012). Valuing one's self: Medial prefrontal involvement in epistemic and emotive investments in self-views. *Cerebral Cortex*, 22(3), 659–667.
- Daubney, M., & Bateman, A. (2015). Mentalization-based therapy (MBT): An overview. *Australasian Psychiatry*, 23(2), 132–135. <https://doi.org/10.1177/1039856214566830>
- Davey, C. G., Allen, N. B., Harrison, B. J., & Yücel, M. (2011). Increased Amygdala Response to Positive Social Feedback in Young People with Major Depressive Disorder. *Biological Psychiatry*, 69(8), 734–741. <https://doi.org/10.1016/j.biopsych.2010.12.004>
- Davidai, S., & Deri, S. (2019). The second pugilist's plight: Why people believe they are above average but are not especially happy about it. *Journal of Experimental Psychology: General*, 148(3), 570.

- Davis, H. (1979). Self-reference and the encoding of personal information in depression. *Cognitive Therapy and Research*, 3, 97-110.
- de Greck, M., Rotte, M., Paus, R., Moritz, D., Thiemann, R., Proesch, U., ... & Northoff, G. (2008). Is our self based on reward? Self-relatedness recruits neural activity in the reward system. *Neuroimage*, 39(4), 2066-2075.
- Delahoy, R., Davey, C. G., Jamieson, A. J., Finlayson-Short, L., Savage, H. S., Steward, T., & Harrison, B. J. (2022). Modulation of the brain's core-self network by self-appraisal processes. *NeuroImage*, 251, 118980. <https://doi.org/10.1016/j.neuroimage.2022.118980>
- Derry, P. A., & Kuiper, N. A. (1981). Schematic processing and self-reference in clinical depression. *Journal of Abnormal Psychology*, 90, 286–297. doi:10.1037/0021-843X.90.4.286
- Dev, A. S., Moore, D. A., Johnson, S. L., & Garrett, K. T. (2022). Sadder≠ wiser: Depressive realism is not robust to replication. *Collabra: Psychology*, 8(1), 38529.
- Dillon, D. G. (2015a). The neuroscience of positive memory deficits in depression. *Frontiers in psychology*, 6, 158950.
- Dillon, D. G., Wiecki, T., Pechtel, P., Webb, C., Goer, F., Murray, L., ... & Pizzagalli, D. A. (2015b). A computational analysis of flanker interference in depression. *Psychological medicine*, 45(11), 2333-2344.
- Dinulescu, S., Alvi, T., Rosenfield, D., Sunahara, C. S., Lee, J., & Tabak, B. A. (2021). Self-Referential Processing Predicts Social Cognitive Ability. *Social Psychological and Personality Science*, 12(1), 99–107. <https://doi.org/10.1177/1948550620902281>
- Disner, S. G., Shumake, J. D., & Beevers, C. G. (2017). Self-referential schemas and attentional bias predict severity and naturalistic course of depression symptoms. *Cognition and Emotion*, 31(4), 632-644.
- Dixon, M. L., Moodie, C. A., Goldin, P. R., Farb, N., Heimberg, R. G., Zhang, J., & Gross, J. J. (2022). Frontoparietal and default mode network contributions to self-referential processing in social anxiety disorder. *Cognitive, Affective, & Behavioral Neuroscience*, 1-12.
- Dufner, M., Gebauer, J. E., Sedikides, C., & Denissen, J. J. (2019). Self-enhancement and psychological adjustment: A meta-analytic review. *Personality and Social Psychology Review*, 23(1), 48-72.
- Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of personality and social psychology*, 57(6), 1082.

- Dunning, D. (1995). Trait importance and modifiability as factors influencing self-assessment and self-enhancement motives. *Personality and Social Psychology Bulletin*, 21(12), 1297-1306.
- Duran, K. A., O'Halloran, H., Soder, H., Yasin, S., Kramer, R., Rosen, S., Brenya, J., Chavarria, K., Savitska, L., & Keenan, J. P. (2021). The medial prefrontal cortex: A potential link between self-deception and affect. *International Journal of Neuroscience*, 131(7), 701–707. <https://doi.org/10.1080/00207454.2020.1753729>
- Eaton, N. R., Krueger, R. F., Keyes, K. M., Wall, M., Hasin, D. S., Markon, K. E., Skodol, A. E., & Grant, B. F. (2013). The Structure and Predictive Validity of the Internalizing Disorders. *Journal of Abnormal Psychology*, 122(1), 86–92. <https://doi.org/10.1037/a0029598>
- Eisenberger, N. I., Inagaki, T. K., Muscatell, K. a, Byrne Haltom, K. E., & Leary, M. R. (2011). The neural sociometer: Brain mechanisms underlying state self-esteem. *Journal of Cognitive Neuroscience*, 23(11), 3448–3455.
- 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.
- Erle, T. M., Barth, N., & Topolinski, S. (2019). Egocentrism in sub-clinical depression. *Cognition and Emotion*, 33(6), 1239–1248. <https://doi.org/10.1080/02699931.2018.1552120>
- Eshel, N., & Roiser, J. P. (2010). Reward and Punishment Processing in Depression. *Biological Psychiatry*, 68(2), 118–124. <https://doi.org/10.1016/j.biopsych.2010.01.027>
- Everaert, J., Podina, I. R., & Koster, E. H. W. (2017). A comprehensive meta-analysis of interpretation biases in depression. *Clinical Psychology Review*, 58, 33–48. <https://doi.org/10.1016/j.cpr.2017.09.005>
- Exline, J. J., & Hill, P. C. (2012). Humility: A consistent and robust predictor of generosity. *The Journal of Positive Psychology*, 7(3), 208-218.
- Feeser, M., Schlagenhauf, F., Sterzer, P., Park, S., Stoy, M., Gutwinski, S., Dalanay, U., Kienast, T., Bauer, M., Heinz, A., Ströhle, A., & Birmahler, F. (2013). Context insensitivity during positive and negative emotional expectancy in depression assessed with functional magnetic resonance imaging. *Psychiatry Research: Neuroimaging*, 212(1), 28–35. <https://doi.org/10.1016/j.psychres.2012.11.010>
- Ferrari, M., Hunt, C., Harrysunker, A., Abbott, M. J., Beath, A. P., & Einstein, D. A. (2019). Self-compassion interventions and psychosocial outcomes: A meta-analysis of RCTs. *Mindfulness*, 10, 1455-1473.
- Festinger L. 1954. A theory of social comparison processes. *Human Relations*.

- Flagan, T., Mumford, J. A., & Beer, J. S. (2017). How Do You See Me? The Neural Basis of Motivated Meta-perception. *Journal of Cognitive Neuroscience*, 29(11), 1908–1917. https://doi.org/10.1162/jocn_a_01169
- Fleming, J. S., & Courtney, B. E. (1984). The dimensionality of self-esteem: II. Hierarchical facet model for revised measurement scales. *Journal of Personality and Social Psychology*, 46(2), 404–421. <https://doi.org/10.1037/0022-3514.46.2.404>
- Fonagy, P., & Bateman, A. W. (2006). Mechanisms of change in mentalization-based treatment of BPD. *Journal of Clinical Psychology*, 62(4), 411–430. <https://doi.org/10.1002/jclp.20241>
- Forbes, E. E. (2009). Where’s the fun in that? Broadening the focus on reward function in depression. *Biological psychiatry*, 66(3), 199-200.
- Forbes, E. E., & Dahl, R. E. (2012). Research Review: altered reward function in adolescent depression: what, when and how? *Journal of Child Psychology and Psychiatry*, 53(1), 3-15.
- Foulkes, L., & Blakemore, S. J. (2016). Is there heightened sensitivity to social reward in adolescence? *Current opinion in neurobiology*, 40, 81-85.
- Freud, A. (1936). *Das Ich und die Abwehrmechanismen*[The Ego and the Mechanisms of Defence]. Wien: Internationaler Psychoanalytischer Verlag.
- Frey, D. L., & Gaertner, S. L. (1986). Helping and the avoidance of inappropriate interracial behavior: A strategy that perpetuates a nonprejudiced self-image. *Journal of Personality and Social Psychology*, 50(6), 1083–1090. <https://doi.org/10.1037/0022-3514.50.6.1083>
- Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry*, 1(2), 148–158. [https://doi.org/10.1016/S2215-0366\(14\)70275-5](https://doi.org/10.1016/S2215-0366(14)70275-5)
- Frith, C. D., & Frith, U. (2006). The Neural Basis of Mentalizing. *Neuron*, 50(4), 531–534. <https://doi.org/10.1016/j.neuron.2006.05.001>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4), 457-472.
- Gilbert, P., & Miles, J. N. V. (2000). Sensitivity to Social Put-Down: It’s relationship to perceptions of social rank, shame, social anxiety, depression, anger and self-other blame. *Personality and Individual Differences*, 29(4), 757–774. [https://doi.org/10.1016/S0191-8869\(99\)00230-5](https://doi.org/10.1016/S0191-8869(99)00230-5)
- Gilboa, E., Roberts, J. E., & Gotlib, I. H. (1997). The effects of induced and naturally occurring dysphoric mood on biases in self-evaluation and memory. *Cognition & Emotion*, 11(1), 65-82.

- Ginot, E. (2012). Self-narratives and dysregulated affective states: The neuropsychological links between self-narratives, attachment, affect, and cognition. *Psychoanalytic Psychology*, 29(1), 59.
- Goldin, P. R., Jazaieri, H., Ziv, M., Kraemer, H., Heimberg, R. G., & Gross, J. J. (2013). Changes in Positive Self-Views Mediate the Effect of Cognitive-Behavioral Therapy for Social Anxiety Disorder. *Clinical Psychological Science*, 1(3), 301–310. <https://doi.org/10.1177/2167702613476867>
- Goldstein, B. L., Hayden, E. P., & Klein, D. N. (2015). Stability of self-referent encoding task performance and associations with change in depressive symptoms from early to middle childhood. *Cognition and Emotion*, 29(8), 1445-1455.
- Gotlib, I. H., & Joormann, J. (2010). Cognition and Depression: Current Status and Future Directions. *Annual Review of Clinical Psychology*, 6(1), 285–312. <https://doi.org/10.1146/annurev.clinpsy.121208.131305>
- Goto, S. G., Lewis, R. S., & Grayzel-Ward, S. (2021). Culture and Self-Construal 15. The Oxford handbook of cultural neuroscience and global mental health, 100, 288.
- Gregg, A. P., Hepper, E. G., & Sedikides, C. (2011). Quantifying self-motives: Functional links between dispositional desires. *European Journal of Social Psychology*, 41(7), 840-852.
- Gregory, B., & Peters, L. (2017). Changes in the self during cognitive behavioural therapy for social anxiety disorder: A systematic review. *Clinical psychology review*, 52, 1-18.
- 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>
- Gruenenfelder-Steiger, A. E., Harris, M. A., & Fend, H. A. (20161003). Subjective and objective peer approval evaluations and self-esteem development: A test of reciprocal, prospective, and long-term effects. *Developmental Psychology*, 52(10), 1563. <https://doi.org/10.1037/dev0000147>
- Grühn, D. (2016). An English word database of emotional terms (EMOTE). *Psychological reports*, 119(1), 290-308.
- Gurel-Atay, E., Xie, G. X., Chen, J., & Kahle, L. R. (2010). Changes in social values in the United States: 1976–2007: “Self-respect” is on the upswing as “a sense of belonging” becomes less important. *Journal of Advertising Research*, 50(1), 57-67.
- Guthrie, T. D., Benadjaoud, Y. Y., & Chavez, R. S. (2022). Social Relationship Strength Modulates the Similarity of Brain-to-Brain Representations of Group Members. *Cerebral Cortex*, 32(11), 2469–2477. <https://doi.org/10.1093/cercor/bhab355>
- Haney, P., & Durlak, J. A. (1998). Changing self-esteem in children and adolescents: A meta-analytical review. *Journal of clinical child psychology*, 27(4), 423-433.

- Hanna, J. T. (2015). Ostracism Increases Positive Valence Theory of Mind Decoding Accuracy.
- Harkness, K. L., Jacobson, J. A., Duong, D., & Sabbagh, M. A. (2010). Mental state decoding in past major depression: Effect of sad versus happy mood induction. *Cognition and Emotion*, 24(3), 497–513. <https://doi.org/10.1080/02699930902750249>
- Harkness, K. L., Washburn, D., Theriault, J. E., Lee, L., & Sabbagh, M. A. (2011). Maternal history of depression is associated with enhanced theory of mind in depressed and nondepressed adult women. *Psychiatry Research*, 189(1), 91–96. <https://doi.org/10.1016/j.psychres.2011.06.007>
- Harrewijn, A., Schmidt, L. A., Westenberg, P. M., Tang, A., & van der Molen, M. J. (2017). Electro cortical measures of information processing biases in social anxiety disorder: a review. *Biological Psychology*, 129, 324–348.
- Harter, S., Stocker, C., & Robinson, N. S. (1996). The perceived directionality of the link between approval and self-worth: The liabilities of a looking gladd self-orientation among young adolescents. *Journal of Research on Adolescence*, 6, 285–308.
- He, J. C., & Côté, S. (2019). Self-insight into emotional and cognitive abilities is not related to higher adjustment. *Nature Human Behaviour*, 3(8), Article 8. <https://doi.org/10.1038/s41562-019-0644-0>
- He, Z., Zhang, D., Muhlert, N., & Elliott, R. (2019). Neural substrates for anticipation and consumption of social and monetary incentives in depression. *Social Cognitive and Affective Neuroscience*, 14(8), 815–826. <https://doi.org/10.1093/scan/nsz061>
- Heatherton, T. F., & Polivy, J. (1991). Development and validation of a scale for measuring state self-esteem. *Journal of Personality and Social Psychology*, 60(6), 895–910. <https://doi.org/10.1037/0022-3514.60.6.895>
- Heeren, A., Dricot, L., Billieux, J., Philippot, P., Grynberg, D., De Timary, P., & Maurage, P. (2017). Correlates of social exclusion in social anxiety disorder: an fMRI study. *Scientific Reports*, 7(1), 260.
- Heerey, E. A., Keltner, D., & Capps, L. M. (2003). Making sense of self-conscious emotion: linking theory of mind and emotion in children with autism. *Emotion*, 3(4), 394.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and brain sciences*, 33(2-3), 111.
- Henriksen, I. O., Ranøyen, I., Indredavik, M. S., & Stenseng, F. (2017). The role of self-esteem in the development of psychiatric problems: A three-year prospective study in a clinical sample of adolescents. *Child and Adolescent Psychiatry and Mental Health*, 11(1), 68. <https://doi.org/10.1186/s13034-017-0207-y>

- Hepper, E. G., Hart, C. M., Gregg, A. P., & Sedikides, C. (2011). Motivated Expectations of Positive Feedback in Social Interactions. *The Journal of Social Psychology*, 151(4), 455–477. <https://doi.org/10.1080/00224545.2010.503722>
- Hitchcock, P. F., Britton, W. B., Mehta, K. P., & Frank, M. J. (2023). Self-judgment dissected: A computational modeling analysis of self-referential processing and its relationship to trait mindfulness facets and depression symptoms. *Cognitive, Affective, & Behavioral Neuroscience*, 23(1), 171-189.
- Hobbs, C., Sui, J., Kessler, D., Munafò, M. R., & Button, K. S. (2021). Self-processing in relation to emotion and reward processing in depression. *Psychological Medicine*, 1–13. <https://doi.org/10.1017/S0033291721003597>
- Hogg, M. A., Sherman, D. K., Dierselhuis, J., Maitner, A. T., & Moffitt, G. (2007). Uncertainty, entitativity, and group identification. *Journal of Experimental Social Psychology*, 43(1), 135–142. <https://doi.org/10.1016/j.jesp.2005.12.008>
- Hohman, Z. P., & Brown, J. K. (2020). Self-Esteem and Self-Enhancement. *Oxford Research Encyclopedia of Psychology*. <https://doi.org/10.1093/acrefore/9780190236557.013.246>
- Hopkins, A. K., Dolan, R., Button, K., & Moutoussis, M. (2019). Reduced positive evidence within activated self-schema may underpin increased sensitivity to negative evaluation in socially anxious individuals. *PsyArXiv*. <https://doi.org/10.31234/osf.io/kf4yz>
- Hopkins, A. K., Dolan, R., Button, K. S., & Moutoussis, M. (2021). A Reduced Self-Positive Belief Underpins Greater Sensitivity to Negative Evaluation in Socially Anxious Individuals. *Computational Psychiatry (Cambridge, Mass.)*, 5(1), 21–37. <https://doi.org/10.5334/cpsy.57>
- Hsu, D. T., Sanford, B. J., Meyers, K. K., Love, T. M., Hazlett, K. E., Walker, S. J., ... & Zubietta, J. K. (2015). It still hurts: altered endogenous opioid activity in the brain during social rejection and acceptance in major depressive disorder. *Molecular psychiatry*, 20(2), 193-200.
- Hsu, D. T., & Jarcho, J. M. (2021). “Next up for psychiatry: Rejection sensitivity and the social brain.” *Neuropsychopharmacology*, 46(1), Article 1. <https://doi.org/10.1038/s41386-020-00802-9>
- Hsu, T.-Y., Liu, T.-L., Cheng, P. Z., Lee, H.-C., Lane, T. J., & Duncan, N. W. (2021). Depressive rumination is correlated with brain responses during self-related processing. *MedRxiv*, 2021.03.18.21253930. <https://doi.org/10.1101/2021.03.18.21253930>
- Hu, X., Kaplan, S., Wei, F., & Vega, R. P. (2014). Do Employees Know How Their Supervisors View Them? A Study Examining Metaperceptions of Job Performance. *Human Performance*, 27(5), 435–457. <https://doi.org/10.1080/08959285.2014.956177>
- Huang, T. C.-C., Hill, C. E., Strauss, N., Heyman, M., & Hussain, M. (2016). Corrective relational experiences in psychodynamic-interpersonal psychotherapy: Antecedents,

- types, and consequences. *Journal of Counseling Psychology*, 63, 183–197.
<https://doi.org/10.1037/cou0000132>
- 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.
- Hughes, B. L., Leong, J. K., Shiv, B., & Zaki, J. (2018). Wanting to like: Motivation influences behavioral and neural responses to social feedback (p. 300657).
<https://doi.org/10.1101/300657>
- Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., Van Zalk, M. H., ... & Back, M. D. (2018). Enhanced versus simply positive: A new condition-based regression analysis to disentangle effects of self-enhancement from effects of positivity of self-view. *Journal of Personality and Social Psychology*, 114(2), 303.
- Ingram, R. E. (1990). Self-focused attention in clinical disorders: review and a conceptual model. *Psychological bulletin*, 107(2), 156.
- International Classification of Diseases, Eleventh Revision (ICD-11), World Health Organization (WHO) 2021
- Ito, A., Yoshida, K., Takeda, K., Sawamura, D., Murakami, Y., Hasegawa, A., Sakai, S., & Izuma, K. (2020). The role of the ventromedial prefrontal cortex in automatic formation of impression and reflected impression. *Human Brain Mapping*, 41(11), 3045–3058.
<https://doi.org/10.1002/hbm.24996>
- James, W. (1890). *The Principles of Psychology* (Vol. 1). New York: Holt.
- Jankowski, K. F., Moore, W. E., Merchant, J. S., Kahn, L. E., & Pfeifer, J. H. (2014). But do you think I'm cool?: Developmental differences in striatal recruitment during direct and reflected social self-evaluations. *Developmental Cognitive Neuroscience*, 8, 40–54.
<https://doi.org/10.1016/j.dcn.2014.01.003>
- Jankowski, K. F., Batres, J., Scott, H., Smyda, G., Pfeifer, J. H., & Quevedo, K. (2018). Feeling left out: depressed adolescents may atypically recruit emotional salience and regulation networks during social exclusion. *Social cognitive and affective neuroscience*, 13(8), 863-876.
- John, O. P., & Robins, R. W. (1993). Determinants of Interjudge Agreement on Personality Traits: The Big Five Domains, Observability, Evaluativeness, and the Unique Perspective of the Self. *Journal of Personality*, 61(4), 521–551. <https://doi.org/10.1111/j.1467-6494.1993.tb00781.x>
- Jones, E. B., & Sharpe, L. (2017). Cognitive bias modification: A review of meta-analyses. *Journal of affective disorders*, 223, 175-183.

- Josephs, R. A., Bosson, J. K., & Jacobs, C. G. (2003). Self-Esteem Maintenance Processes: Why Low Self-Esteem may be Resistant to Change. *Personality and Social Psychology Bulletin*, 29(7), 920–933. <https://doi.org/10.1177/0146167203029007010>
- Kaplan, H. B. (1986). *Social psychology of self-referent behavior*. Springer Science & Business Media.
- Kaplan, S. A., Santuzzi, A. M., & Ruscher, J. B. (2009). Elaborative Metaperceptions in Outcome-Dependent Situations: The Diluted Relationship Between Default Self-Perceptions and Metaperceptions. *Social Cognition*, 27(4), 601–614. <https://doi.org/10.1521/soco.2009.27.4.601>
- Karpen, S. C. (2018). The social psychology of biased self-assessment. *American Journal of Pharmaceutical Education*, 82(5), 6299.
- Kawamoto, T., Nittono, H., & Ura, M. (2014). Social exclusion induces early-stage perceptual and behavioral changes in response to social cues. *Social Neuroscience*, 9(2), 174–185. <https://doi.org/10.1080/17470919.2014.883325>
- Ke, T., Wu, J., Willner, C. J., Brown, Z., Banz, B., Van Noordt, S., Waters, A. C., & Crowley, M. J. (2020). The glass is half empty: Negative self-appraisal bias and attenuated neural response to positive self-judgment in adolescence. *Social Neuroscience*, 15(2), 140–157. <https://doi.org/10.1080/17470919.2019.1697744>
- Keane, L., & Loades, M. (2017). Low self-esteem and internalizing disorders in young people—a systematic review. *Child and Adolescent Mental Health*, 22(1), 4-15.
- 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.
- Kendler, K. S., Prescott, C. A., Myers, J., & Neale, M. C. (2003). The Structure of Genetic and Environmental Risk Factors for Common Psychiatric and Substance Use Disorders in Men and Women. *Archives of General Psychiatry*, 60(9), 929–937. <https://doi.org/10.1001/archpsyc.60.9.929>
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis* (pp. xvii, 270). Guilford Press.
- Kernis, M. H., & Goldman, B. M. (2003). Stability and variability in self-concept and self-esteem. In *Handbook of self and identity* (pp. 106–127). The Guilford Press.
- Kernis, M. H., Grannemann, B. D., & Barclay, L. C. (1989). Stability and level of self-esteem as predictors of anger arousal and hostility. *Journal of Personality and Social Psychology*, 56, 1013–1022. <https://doi.org/10.1037/0022-3514.56.6.1013>
- Kernis M. H., Whisenhunt Connie R., Waschull Stefanie B., Greenier Keegan D., Berry Andrea J., Herlocker Caryn E., & Anderson Craig A. (1998). Multiple Facets of Self-Esteem and

- their Relations to Depressive Symptoms. *Personality and Social Psychology Bulletin*, 24(6). <https://doi.org/10.1177/0146167298246009>
- Kiang M., Farzan F., Blumberger D. M., Kutas M., McKinnon M. C., Kansal V., Rajji T. K., Daskalakis Z. J. (2017). Abnormal self-schema in semantic memory in major depressive disorder: Evidence from event-related brain potentials. *Biological Psychiatry*, 126, 41–47. <https://doi.org/10.1016/j.biopsycho.2017.04.003>
- Kim, H., Di Domenico, S. I., & Connelly, B. S. (2019). Self–Other Agreement in Personality Reports: A Meta-Analytic Comparison of Self- and Informant-Report Means. *Psychological Science*, 30(1), 129–138. <https://doi.org/10.1177/0956797618810000>
- Kim, H., Lee, H., Lo, R. F., Suh, E. M., & Schimmack, U. (2022). Seeing the self through rose-colored glasses: A cross-cultural study of positive illusions using a behavioral approach. *PloS one*, 17(10), e0274535.
- Kim, Y.-H., Chiu, C., & Zou, Z. (2010). Know thyself: Misperceptions of actual performance undermine achievement motivation, future performance, and subjective well-being. *Journal of Personality and Social Psychology*, 99, 395–409. <https://doi.org/10.1037/a0020555>
- Kinch, J. W. (1964). A Formalized Theory of the Self-Concept | *American Journal of Sociology*: Vol 68, No 4. <https://www.journals.uchicago.edu/doi/abs/10.1086/223404>
- Kircanski, K., Mazur, H., & Gotlib, I. H. (2013). Behavioral activation system moderates self-referent processing following recovery from depression. *Psychological Medicine*, 43(9), 1909–1919. <https://doi.org/10.1017/S0033291712002851>
- Kircanski, K., Thompson, R. J., Sorenson, J. E., Sherdell, L., & Gotlib, I. H. (2015). Rumination and Worry in Daily Life: Examining the Naturalistic Validity of Theoretical Constructs. *Clinical Psychological Science*, 3(6), 926–939. <https://doi.org/10.1177/2167702614566603>
- Kitayama, S., & Uchida, Y. (2003). Explicit self-criticism and implicit self-regard: Evaluating self and friend in two cultures. *Journal of Experimental Social Psychology*, 39(5), 476–482.
- Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., Wager, T. D., & Arch, J. J. (2017). Social anxiety is characterized by biased learning about performance and the self. *Emotion*, 17(8), 1144–1155. <https://doi.org/10.1037/emo0000296>
- Komulainen, E., Heikkilä, R., Nummenmaa, L., Raji, T. T., Harmer, C. J., Isometsä, E., & Ekelund, J. (2018). Short-term escitalopram treatment normalizes aberrant self-referential processing in major depressive disorder. *Journal of Affective Disorders*, 236, 222–229. <https://doi.org/10.1016/j.jad.2018.04.096>

- Koole, S. L., Schwager, S., & Rothermund, K. (2015). Resilience is more about being flexible than about staying positive. *Behavioral and Brain Sciences*, 38.
- Korrelboom, K., IJdema, T., Karreman, A., & van der Gaag, M. (2022). The effectiveness of transdiagnostic applications of competitive memory training (COMET) on low self-esteem and comorbid depression: A meta-analysis of randomized controlled trials. *Cognitive Therapy and Research*, 46(3), 532-543.
- Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively Biased Processing of Self-Relevant Social Feedback. *Journal of Neuroscience*, 32(47), 16832–16844. <https://doi.org/10.1523/JNEUROSCI.3016-12.2012>
- Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., ... & Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of abnormal psychology*, 126(4), 454.
- Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. <https://doi.org/10.3389/neuro.06.004.2008>
- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(5), Article 5. <https://doi.org/10.1038/nn.2303>
- Krueger, J. (1998). Enhancement Bias in Description of Self and Others.
- Kube, T., Kirchner, L., Rief, W., Gärtner, T., & Glombiewski, J. A. (2019). Belief updating in depression is not related to increased sensitivity to unexpectedly negative information. *Behaviour Research and Therapy*, 123, 103509. <https://doi.org/10.1016/j.brat.2019.103509>
- Kube, T., & Rozenkrantz, L. (2021). When Beliefs Face Reality: An Integrative Review of Belief Updating in Mental Health and Illness. *Perspectives on Psychological Science*, 16(2), 247–274. <https://doi.org/10.1177/1745691620931496>
- Kuck, N., Cafitz, L., Bürkner, P.-C., Hoppen, L., Wilhelm, S., & Buhlmann, U. (2021). Body dysmorphic disorder and self-esteem: A meta-analysis. *BMC Psychiatry*, 21(1), 310. <https://doi.org/10.1186/s12888-021-03185-3>
- Kuckertz, J. M., & Amir, N. (2017). Cognitive bias modification. In *The science of cognitive behavioral therapy* (pp. 463-491). Academic Press.
- Kuiper, N. A., & Rogers, T. B. (1979). Encoding of personal information: Self–other differences. *Journal of Personality and Social Psychology*, 37(4), 499.

- Kuiper, N. A., & Derry, P. A. (1982). Depressed and nondepressed content self-reference in mild depressives. *Journal of Personality*, 50(1), 67–80. <https://doi.org/10.1111/j.1467-6494.1982.tb00746.x>
- Kujawa, A., & Burkhouse, K. L. (2017). Vulnerability to depression in youth: Advances from affective neuroscience. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2(1), 28–37.
- Kumar, P., Waiter, G. D., Dubois, M., Milders, M., Reid, I., & Steele, J. D. (2017). Increased neural response to social rejection in major depression. *Depression and anxiety*, 34(11), 1049-1056.
- Kurt, A., & Paulhus, D. L. (2008). Moderators of the adaptiveness of self-enhancement: Operationalization, motivational domain, adjustment facet, and evaluator. *Journal of Research in Personality*, 42(4), 839–853. <https://doi.org/10.1016/j.jrp.2007.11.005>
- LaBouff, J. P., Rowatt, W. C., Johnson, M. K., Tsang, J. A., & Willerton, G. M. (2012). Humble persons are more helpful than less humble persons: Evidence from three studies. *The Journal of Positive Psychology*, 7(1), 16-29
- Leary, M. M. R., Tambor, E. S. E., Terdal, S. K., & Downs, D. L. (1995). Self-esteem as an interpersonal monitor: The sociometer hypothesis. *Journal of Personality and Social Psychology*, 68(3), 518–530. <https://doi.org/10.1037//0022-3514.68.3.518>
- Leary, M. R., & Baumeister, R. F. (2000). The nature and function of self-esteem: Sociometer theory. In *Advances in Experimental Social Psychology* (Vol. 32, pp. 1–62). Academic Press. [https://doi.org/10.1016/S0065-2601\(00\)80003-9](https://doi.org/10.1016/S0065-2601(00)80003-9)
- Leary, M. R., & MacDonald, G. (2003). Individual differences in self-esteem: A review and theoretical integration.
- Leary, M. R. (2007). *The curse of the self: Self-awareness, egotism, and the quality of human life*. Oxford University Press.
- Lebeau, R. T., Glenn, D. E., Hanover, L. N., Beesdo-Baum, K., Wittchen, H. U., & Craske, M. G. (2012). A dimensional approach to measuring anxiety for DSM-5. *International Journal of Methods in Psychiatric Research*, 21(4), 258-272.
- LeBeau, R. T., Mesri, B., & Craske, M. G. (2016). The DSM-5 social anxiety disorder severity scale: Evidence of validity and reliability in a clinical sample. *Psychiatry research*, 244, 94-96.
- Leckelt, M., Küfner, A. C., Nestler, S., & Back, M. D. (2015). Behavioral processes underlying the decline of narcissists' popularity over time. *Journal of Personality and Social Psychology*, 109(5), 856.

- Lee, L., Harkness, K. L., Sabbagh, M. A., & Jacobson, J. A. (2005). Mental state decoding abilities in clinical depression. *Journal of Affective Disorders*, 86(2–3), 247–258. <https://doi.org/10.1016/j.jad.2005.02.007>
- Lee, E. J., & Stone, S. I. (2012). Co-occurring internalizing and externalizing behavioral problems: The mediating effect of negative self-concept. *Journal of youth and adolescence*, 41, 717-731.
- Legault, L., Coleman, D., Jurchak, K., & Scaltsas, N. (2022). Reducing prejudice by enhancing the other rather than the self. *Self and Identity*, 21(6), 687-709.
- Leitner, J. B., Hehman, E., Jones, J. M., & Forbes, C. E. (2014). Self-enhancement Influences Medial Frontal Cortex Alpha Power to Social Rejection Feedback. http://www.mitpressjournals.org/doi/abs/10.1162/jocn_a_00645#.VEqNrovF-hQ
- Lemogne, C., le Bastard, G., Mayberg, H., Volle, E., Bergouignan, L., Lehericy, S., Allilaire, J.-F., & Fossati, P. (2009). In search of the depressive self: Extended medial prefrontal network during self-referential processing in major depression. *Social Cognitive and Affective Neuroscience*, 4(3), 305–312. <https://doi.org/10.1093/scan/nsp008>
- Lemogne, C., Delaveau, P., Freton, M., Guionnet, S., & Fossati, P. (2012). Medial prefrontal cortex and the self in major depression. *Journal of Affective Disorders*, 136(1), e1–e11. <https://doi.org/10.1016/j.jad.2010.11.034>
- LeMoult, J., Yoon, K. L., & Joormann, J. (2016). Rumination and cognitive distraction in major depressive disorder: An examination of respiratory sinus arrhythmia. *Journal of psychopathology and behavioral assessment*, 38, 20-29.
- LeMoult, J., Kircanski, K., Prasad, G., & Gotlib, I. H. (2017). Negative self-referential processing predicts the recurrence of major depressive episodes. *Clinical Psychological Science*, 5(1), 174-181.
- LeMoult, J., & Gotlib, I. H. (2019). Depression: A cognitive perspective. *Clinical psychology review*, 69, 51-66.
- Letzring, T. D., & Funder, D. C. (2018). Interpersonal accuracy in trait judgments. In *The SAGE handbook of personality and individual differences: Applications of personality and individual differences* (pp. 253–282). Sage Reference. <https://doi.org/10.4135/9781526451248.n11>
- Li, H., Zeigler-Hill, V., Yang, J., Jia, L., Xiao, X., Luo, J., & Zhang, Q. (2012). Low self-esteem and the neural basis of attentional bias for social rejection cues: Evidence from the N2pc ERP component. *Personality and Individual Differences*, 53(8), 947–951. <https://doi.org/10.1016/j.paid.2012.03.004>
- Li, Y., Li, M., Wei, D., Kong, X., Du, X., Hou, X., Sun, J., & Qiu, J. (2017). Self-referential processing in unipolar depression: Distinct roles of subregions of the medial prefrontal

- cortex. *Psychiatry Research: Neuroimaging*, 263, 8–14.
<https://doi.org/10.1016/j.psychresns.2017.02.008>
- Liknaitzky, P., Smillie, L. D., & Allen, N. B. (2017). Out-of-the-Blue: Depressive Symptoms are Associated with Deficits in Processing Inferential Expectancy-Violations Using a Novel Cognitive Rigidity Task. *Cognitive Therapy and Research*, 41(5), 757–776.
<https://doi.org/10.1007/s10608-017-9853-x>
- Lieberman, M. D., & Pfeifer, J. H. (2004). The self and social perception: Three kinds of questions in social cognitive neuroscience. In *The cognitive neuroscience of social behaviour* (pp. 207-248). Psychology Press.
- Lin, Y., Callahan, C. P., & Moser, J. S. (2018). A mind full of self: Self-referential processing as a mechanism underlying the therapeutic effects of mindfulness training on internalizing disorders. *Neuroscience & Biobehavioral Reviews*, 92, 172–186.
<https://doi.org/10.1016/j.neubiorev.2018.06.007>
- Loeb, E., Hessel, E. T., & Allen, J. (2016). The self-fulfilling prophecy of adolescent social expectations. *International Journal of Behavioral Development*, 40(6), 555–564.
<https://doi.org/10.1177/0165025415618274>
- Loeb, E. L., Tan, J. S., Hessel, E. T., & Allen, J. P. (2018). Getting What You Expect: Negative Social Expectations in Early Adolescence Predict Hostile Romantic Partnerships and Friendships Into Adulthood. *The Journal of Early Adolescence*, 38(4), 475–496.
<https://doi.org/10.1177/0272431616675971>
- Low, A. A. Y., Hopper, W. J. T., Angelescu, I., Mason, L., Will, G.-J., & Moutoussis, M. (2021). Self-Esteem depends on beliefs about the rate of change of social approval. *PsyArXiv*.
<https://doi.org/10.31234/osf.io/rhvdu>
- Luan, Z., Poorthuis, A. M. G., Hutteman, R., Asendorpf, J. B., Denissen, J. J. A., & van Aken, M. A. G. (2018). See me through my eyes: Adolescent–parent agreement in personality predicts later self-esteem development. *International Journal of Behavioral Development*, 42(1), 17–25. <https://doi.org/10.1177/0165025417690263>
- Luber, B., Lou, H. C., Keenan, J. P., & Lisanby, S. H. (2012). Self-enhancement processing in the default network: A single-pulse TMS study. *Experimental Brain Research*, 223(2), 177–187. <https://doi.org/10.1007/s00221-012-3249-7>
- Luking, K. R., Pagliaccio, D., Luby, J. L., & Barch, D. M. (2016). Reward processing and risk for depression across development. *Trends in cognitive sciences*, 20(6), 456-468.
- Luyten, P., & Fonagy, P. (2015). The neurobiology of mentalizing. *Personality Disorders: Theory, Research, and Treatment*, 6, 366–379. <https://doi.org/10.1037/per0000117>
- Luyten, P., & Fonagy, P. (2018). The stress–reward–mentalizing model of depression: An integrative developmental cascade approach to child and adolescent depressive disorder

- based on the Research Domain Criteria (RDoC) approach. *Clinical Psychology Review*, 64, 87–98. <https://doi.org/10.1016/j.cpr.2017.09.008>
- Lynch, B. P., & vanDellen, M. R. (2020). A multi-motive framework for predicting variability in self-enhancement. *Social and Personality Psychology Compass*, 14(10), e12561. <https://doi.org/10.1111/spc3.12561>
- MacDonald, M. R., & Kuiper, N. A. (1985). Efficiency and automaticity of self-schema processing in clinical depressives. *Motivation and Emotion*, 9(2), 171-184.
- Malda-Castillo, J., Browne, C., & Perez-Algorta, G. (2019). Mentalization-based treatment and its evidence-base status: A systematic literature review. *Psychology and Psychotherapy*, 92(4), 465–498. <https://doi.org/10.1111/papt.12195>
- Maleki, G., Zabihzadeh, A., Richman, M. J., Demetrovics, Z., & Mohammadnejad, F. (2020). Decoding and reasoning mental states in major depression and social anxiety disorder. *BMC Psychiatry*, 20(1), 463. <https://doi.org/10.1186/s12888-020-02873-w>
- Mattek, A. M., Wolford, G. L., & Whalen, P. J. (2017). A mathematical model captures the structure of subjective affect. *Perspectives on Psychological Science*, 12(3), 508-526.
- Markowitz, J. C., Milrod, B., Luyten, P., & Holmqvist, R. (2019). Mentalizing in Interpersonal Psychotherapy. *American Journal of Psychotherapy*, 72(4), 95–100. <https://doi.org/10.1176/appi.psychotherapy.20190021>
- Markus, H. R., & Kitayama, S. (1991). Cultural variation in the self-concept. In *The self: Interdisciplinary approaches* (pp. 18-48). New York, NY: Springer New York.
- Martinsen, K. D., Rasmussen, L.-M. P., Wentzel-Larsen, T., Holen, S., Sund, A. M., Pedersen, M. L., Løvaas, M. E. S., Patras, J., Adolfsen, F., & Neumer, S.-P. (2021). Change in quality of life and self-esteem in a randomized controlled CBT study for anxious and sad children: Can targeting anxious and depressive symptoms improve functional domains in schoolchildren? *BMC Psychology*, 9(1), 8. <https://doi.org/10.1186/s40359-021-00511-y>
- McCarthy, P. A., & Morina, N. (2020). Exploring the association of social comparison with depression and anxiety: A systematic review and meta-analysis. *Clinical psychology & psychotherapy*, 27(5), 640-671.
- McCrae, R. R., & Costa, P. T. (1987). Validation of the Five-Factor Model of Personality Across Instruments and Observers. *Journal of Personality and Social Psychology*, 52(1), 81–90. <https://doi.org/10.1037/0022-3514.52.1.81>
- McEvoy, P. M., Hyett, M. P., Shihata, S., Price, J. E., & Strachan, L. (2019). The impact of methodological and measurement factors on transdiagnostic associations with intolerance of uncertainty: A meta-analysis. *Clinical Psychology Review*, 73, 101778. <https://doi.org/10.1016/j.cpr.2019.101778>

- McKay, R. T., & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, 32(6), 493-510.
- Mead, G. H. (1934). *Mind, Self, and Society from the Standpoint of a Social Behaviorist*. Chicago.
- Mecca, A., Smelser, N. J., & Vasconcellos, J. (Eds.). (1989). *The social importance of self-esteem*. Univ of California Press.
- Mellick, W., & Sharp, C. (2016). Mental State Decoding in Adolescent Boys with Major Depressive Disorder versus Sex-Matched Healthy Controls. *Psychopathology*, 49(1), 53–59. <https://doi.org/10.1159/000443989>
- Meulen, M. van der, IJzendoorn, M. H. van, & Crone, E. A. (2016). Neural Correlates of Prosocial Behavior: Compensating Social Exclusion in a Four-Player Cyberball Game. *PLOS ONE*, 11(7), e0159045. <https://doi.org/10.1371/journal.pone.0159045>
- Miers, A. C. (2021). An Investigation Into the Influence of Positive Peer Feedback on Self-Relevant Cognitions in Social Anxiety. *Behaviour Change*, 38(3), 193–207. <https://doi.org/10.1017/bec.2021.8>
- Mitchell, J. P., Schirmer, J., Ames, D. L., & Gilbert, D. T. (2011). Medial prefrontal cortex predicts intertemporal choice. *Journal of Cognitive Neuroscience*, 23(4), 857–866. <https://doi.org/10.1162/jocn.2010.21479>
- Moore, M. T., & Fresco, D. M. (2012). Depressive realism: A meta-analytic review. *Clinical Psychology Review*, 32(6), 496–509. <https://doi.org/10.1016/j.cpr.2012.05.004>
- Mor, N., & Winquist, J. (2002). Self-focused attention and negative affect: A meta-analysis. *Psychological Bulletin*, 128(4), 638–662. <https://doi.org/10.1037/0033-2909.128.4.638>
- Moran, E. K., Mehta, N., & Kring, A. M. (2012). Emotional responding in depression: Distinctions in the time course of emotion. *Cognition and Emotion*, 26(7), 1153–1175. <https://doi.org/10.1080/02699931.2011.638909>
- Moritz, D., & Roberts, J. E. (2020). Depressive Symptoms and Self-Esteem as Moderators of Metaperceptions of Social Rejection Versus Acceptance: A Truth and Bias Analysis. *Clinical Psychological Science*, 8(2), 252–265. <https://doi.org/10.1177/2167702619894906>
- Mosch, A., & Borkenau, P. (2016). Psychologically Adjusted Persons Are Less Aware of How They Are Perceived by Others. *Personality and Social Psychology Bulletin*, 42(7), 910–922. <https://doi.org/10.1177/0146167216647383>
- Moscovitch, D. A., Orr, E., Rowa, K., Reimer, S. G., & Antony, M. M. (2009). In the absence of rose-colored glasses: Ratings of self-attributes and their differential certainty and importance across multiple dimensions in social phobia. *Behaviour Research and Therapy*, 47(1), 66-70.

- Nejad, A. B., Fossati, P., & Lemogne, C. (2013). Self-Referential Processing, Rumination, and Cortical Midline Structures in Major Depression. *Frontiers in Human Neuroscience*, 7. <https://doi.org/10.3389/fnhum.2013.00666>
- Nejati, V., Zabihzadeh, A., Nikfarjam, M. R., Pournaghdali, A., Naderi, Z., & Riyahi, M. T. (2012). Self as a social cue: Evidence for reading mind from eyes test. *Procedia - Social and Behavioral Sciences*, 32, 82–88. <https://doi.org/10.1016/j.sbspro.2012.01.015>
- Nelson, B. D., Perlman, G., Klein, D. N., Kotov, R., & Hajcak, G. (2016). Blunted neural response to rewards as a prospective predictor of the development of depression in adolescent girls. *American Journal of Psychiatry*, 173(12), 1223-1230.
- Nestor, B. A., Sutherland, S., & Garber, J. (2022). Theory of mind performance in depression: A meta-analysis. *Journal of Affective Disorders*, 303, 233–244. <https://doi.org/10.1016/j.jad.2022.02.028>
- Ng, T. H., Alloy, L. B., & Smith, D. V. (2019). Meta-analysis of reward processing in major depressive disorder reveals distinct abnormalities within the reward circuit. *Translational Psychiatry*, 9(1), Article 1. <https://doi.org/10.1038/s41398-019-0644-x>
- Ngo, H., VanderLaan, D. P., & Aitken, M. (2020). Self-esteem, symptom severity, and treatment response in adolescents with internalizing problems. *Journal of Affective Disorders*, 273, 183–191. <https://doi.org/10.1016/j.jad.2020.04.045>
- Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on psychological science*, 3(5), 400-424.
- Norem, J. K., & Cantor, N. (1986). Defensive pessimism: Harnessing anxiety as motivation. *Journal of Personality and Social Psychology*, 51, 1208–1217. <https://doi.org/10.1037/0022-3514.51.6.1208>
- Northoff, G., & Hayes, D. J. (2011). Is Our Self Nothing but Reward? *Biological Psychiatry*, 69(11), 1019–1025. <https://doi.org/10.1016/j.biopsych.2010.12.014>
- Northoff, G., & Tumati, S. (2019). “Average is good, extremes are bad”—Non-linear inverted U-shaped relationship between neural mechanisms and functionality of mental features. *Neuroscience & Biobehavioral Reviews*, 104, 11–25.
- O’Callaghan, G., & Stringaris, A. (2019). Reward Processing in Adolescents with Depression. *Cognitive Dimensions of Major Depressive Disorder*, 129.
- O’Connor, L., Berry, J., Lewis, T., Mulherin, K., & Crisostomo, P. (2007). Empathy and depression: The moral system on overdrive. In *Empathy in Mental Illness*. Cambridge University Press.
- O’Mara, E. M., & Gaertner, L. (2017). Does self-enhancement facilitate task performance?. *Journal of Experimental Psychology: General*, 146(3), 442.

- Orchard, F., & Reynolds, S. (2018). The combined influence of cognitions in adolescent depression: Biases of interpretation, self-evaluation, and memory. *British Journal of Clinical Psychology*, 57(4), 420-435.
- Orth, U., & Robins, R. W. (2022). Is high self-esteem beneficial? Revisiting a classic question. *American Psychologist*, 77(1), 5.
- 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>
- Ottenbreit, N. D., & Dobson, K. S. (2004). Avoidance and depression: The construction of the Cognitive–Behavioral Avoidance Scale. *Behaviour Research and Therapy*, 42(3), 293–313. [https://doi.org/10.1016/S0005-7967\(03\)00140-2](https://doi.org/10.1016/S0005-7967(03)00140-2)
- Overholser, J. C., Adams, D. M., Lehnert, K. L., & Brinkman, D. C. (1995). Self-Esteem Deficits and Suicidal Tendencies among Adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 34(7), 919–928. <https://doi.org/10.1097/00004583-199507000-00016>
- Paulus, M. P., Feinstein, J. S., & Khalsa, S. S. (2019). An Active Inference Approach to Interoceptive Psychopathology. *Annual Review of Clinical Psychology*, 15(1), 97–122. <https://doi.org/10.1146/annurev-clinpsy-050718-095617>
- Pe, M. L., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion*, 13(4), 739.
- Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1–2), 8–13. <https://doi.org/10.1016/j.jneumeth.2006.11.017>
- Perinelli, E., Alessandri, G., Cepale, G., & Fraccaroli, F. (2022). The sociometer theory at work: Exploring the organizational interpersonal roots of self-esteem. *Applied Psychology*, 71(1), 76–102. <https://doi.org/10.1111/apps.12312>
- Petereit, P., Jessen, S., Goregliad Fjaellingsdal, T., & Krämer, U. M. (2022). Social Context and Rejection Expectations Modulate Neural and Behavioral Responses to Social Feedback. *Journal of Cognitive Neuroscience*, 34(5), 823–845. https://doi.org/10.1162/jocn_a_01829
- Pfeifer, J. H., Masten, C. L., Borofsky, L. A., Dapretto, M., Fuligni, A. J., & Lieberman, M. D. (2009). Neural correlates of direct and reflected self-appraisals in adolescents and adults: When social perspective-taking informs self-perception. *Child Development*, 80(4), 1016–1038.

- Pfeifer, J. H., & Peake, S. J. (2012). Self-development: Integrating cognitive, socioemotional, and neuroimaging perspectives. *Developmental Cognitive Neuroscience*, 2(1), 55–69. <https://doi.org/10.1016/j.dcn.2011.07.012>
- Pfeifer, J. H., & Allen, N. B. (2021). Puberty Initiates Cascading Relationships Between Neurodevelopmental, Social, and Internalizing Processes Across Adolescence. *Biological Psychiatry*, 89(2), 99–108. <https://doi.org/10.1016/j.biopsych.2020.09.002>
- Phillips, W. J., Hine, D. W., & Thorsteinsson, E. B. (2010). Implicit cognition and depression: A meta-analysis. *Clinical Psychology Review*, 30(6), 691-709.
- Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. *Assessment*, 18(3), 263-283.
- Pitliya, R. J., Nelson, B. D., Hajcak, G., & Jin, J. (2022). Drift-Diffusion Model Reveals Impaired Reward-Based Perceptual Decision-Making Processes Associated with Depression in Late Childhood and Early Adolescent Girls. *Research on Child and Adolescent Psychopathology*, 50(11), 1515-1528.
- Poland, J., Von Eckardt, B., & Spaulding, W. (1994). Problems with the DSM approach to classifying psychopathology. *Philosophical psychopathology*, 235-260.
- Pool, E. R., Tord, D. M., Delplanque, S., Stussi, Y., Cereghetti, D., Vuilleumier, P., & Sander, D. (2022). Differential contributions of ventral striatum subregions to the motivational and hedonic components of the affective processing of reward. *Journal of Neuroscience*, 42(13), 2716-2728.
- Poland, J., & Frank, M. (2017). The promise of computational psychiatry.
- Porter, A. C., Zelkowitz, R. L., Gist, D. C., & Cole, D. A. (2019). Self-evaluation and depressive symptoms: A latent variable analysis of self-esteem, shame-proneness, and self-criticism. *Journal of psychopathology and behavioral assessment*, 41, 257-270.
- Preuss, G. S., & Alicke, M. D. (2017). My worst faults and misdeeds: Self-criticism and self-enhancement can co-exist. *Self and Identity*, 16(6), 645-663.
- Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, 52(4), 449-459.
- Pyszczynski, T., & Greenberg, J. (1985). Depression and preference for self-focusing stimuli after success and failure. *Journal of Personality and Social Psychology*, 49(4), 1066–1075.
- Quevedo, K., Ng, R., Scott, H., Smyda, G., Pfeifer, J. H., & Malone, S. (2017). The neurobiology of self-processing in abused depressed adolescents. *Development and Psychopathology*, 29(3), 1057–1073.

- Rantanen, M., Hautala, J., Loberg, O., Nuorva, J., Hietanen, J. K., Nummenmaa, L., & Astikainen, P. (2021). Attentional bias towards interpersonal aggression in depression—An eye movement study. *Scandinavian journal of psychology*, 62(5), 639-647.
- Rappaport, B. I., & Barch, D. M. (2020). Brain responses to social feedback in internalizing disorders: A comprehensive review. *Neuroscience & Biobehavioral Reviews*, 118, 784–808. <https://doi.org/10.1016/j.neubiorev.2020.09.012>
- Ratcliff, R., & Mc Koon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice.
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in cognitive sciences*, 20(4), 260-281.
- Reichenberger, J., Wiggert, N., Agroskin, D., Wilhelm, F. H., & Blechert, J. (2017). No praise, please: Depressive symptoms, reactivity to positive social interaction, and fear of positive evaluation. *Journal of Behavior Therapy and Experimental Psychiatry*, 54, 186-194.
- Rice, K., Schutte, N. S., Rock, A. J., & Murray, C. V. (2021). Structure, validity and cut-off scores for the APA emerging measure: DSM-5 social anxiety disorder severity scale (SAD-D). *J Depress Anxiety*, 10(406), 2167-1044.
- Robins, R. W., & Beer, J. S. (2001). Positive illusions about the self: short-term benefits and long-term costs. *Journal of personality and social psychology*, 80(2), 340.
- Robins, R. W., & Trzesniewski, K. H. (2005). Self-esteem development across the lifespan. *Current directions in psychological science*, 14(3), 158-162.
- Rodman, A. M., Powers, K. E., & Somerville, L. H. (2017). Development of self-protective biases in response to social evaluative feedback. *Proceedings of the National Academy of Sciences*, 114(50), 13158–13163.
- Roiser, J. P., Elliott, R., & Sahakian, B. J. (2012). Cognitive mechanisms of treatment in depression. *Neuropsychopharmacology*, 37(1), 117-136.
- Rosenberg, M., Schooler, C., Schoenbach, C., & Rosenberg, F. (1995). Global self-esteem and specific self-esteem: Different concepts, different outcomes. *American sociological review*, 141-156.
- Rouault, M., Will, G.-J., Fleming, S., & Dolan, R. (2021). Low self-esteem and the formation of global self-performance estimates in emerging adulthood. *PsyArXiv*. <https://doi.org/10.31234/osf.io/d8ns2>
- Rueger, S. Y., Malecki, C. K., & Demaray, M. K. (2010). Relationship Between Multiple Sources of Perceived Social Support and Psychological and Academic Adjustment in Early Adolescence: Comparisons Across Gender. *Journal of Youth and Adolescence*, 39(1), 47–61. <https://doi.org/10.1007/s10964-008-9368-6>

- Ruscio, A. M. (2019). Normal versus pathological mood: Implications for diagnosis. *Annual review of clinical psychology*, 15, 179-205.
- Sahi, R. S., & Eisenberger, N. I. (2021). Why Don't You Like Me? The Role of the Mentalizing Network in Social Rejection. In M. Gilead & K. N. Ochsner (Eds.), *The Neural Basis of Mentalizing* (pp. 613–628). Springer International Publishing.
https://doi.org/10.1007/978-3-030-51890-5_32
- Schreiter, S., Pijnenborg, G. H. M., & aan het Rot, M. (2013). Empathy in adults with clinical or subclinical depressive symptoms. *Journal of Affective Disorders*, 150(1), 1–16.
<https://doi.org/10.1016/j.jad.2013.03.009>
- Schröder-Abé, M., Rudolph, A., Wiesner, A., & Schütz, A. (2007). Self-esteem discrepancies and defensive reactions to social feedback. *International Journal of Psychology*, 42(3), 174–183. <https://doi.org/10.1080/00207590601068134>
- Schultz, L. T., & Heimberg, R. G. (2008). Attentional focus in social anxiety disorder: Potential for interactive processes. *Clinical Psychology Review*, 28(7), 1206–1221.
<https://doi.org/10.1016/j.cpr.2008.04.003>
- Schütz, A., & Baumeister, R. F. (2017). Positive illusions and the happy mind. *The happy mind: Cognitive contributions to well-being*, 177-193
- Sedikides, C., & Skowronski, J. J. (1995). On the sources of self-knowledge: The perceived primacy of self-reflection. *Journal of Social and Clinical Psychology*, 14, 244–270.
<https://doi.org/10.1521/jscp.1995.14.3.244>
- Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural self-enhancement. *Journal of personality and social psychology*, 84(1), 60.
- Sedikides, C., Gaertner, L., & Vevea, J. L. (2005). Pancultural self-enhancement reloaded: a meta-analytic reply to Heine (2005).
- Sedikides, C., & Skowronski, J. J. (2009). Social cognition and self-cognition: Two sides of the same evolutionary coin?. *European Journal of Social Psychology*, 39(7), 1245-1249.
- Sedikides, C. (2021). In search of narcissus. *Trends in Cognitive Sciences*, 25(1), 67-80.
- Shany, O., Gurevitch, G., Gilam, G., Dunskey, N., Reznik Balter, S., Greental, A., Nutkevitch, N., Eldar, E., & Hendler, T. (2022). A corticostriatal pathway mediating self-efficacy enhancement. *Npj Mental Health Research*, 1(1), Article 1.
<https://doi.org/10.1038/s44184-022-00006-7>
- Shapka, J. D., & Keating, D. P. (2005). Structure and Change in Self-Concept During Adolescence. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 37(2), 83.

- Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, 14(11), Article 11. <https://doi.org/10.1038/nn.2949>
- Shestiyuk, A. Y., & Deldin, P. J. (2010). Automatic and strategic representation of the self in major depression: trait and state abnormalities. *American Journal of Psychiatry*, 167(5), 536-544.
- Shi, Y., Sedikides, C., Cai, H., Liu, Y., & Yang, Z. (2017). Disowning the self: The cultural value of modesty can attenuate self-positivity. *Quarterly Journal of Experimental Psychology*, 70(6), 1023-1032.
- Shin, J. E., Suh, E. M., Li, N. P., Eo, K., Chong, S. C., & Tsai, M. H. (2019). Darling, get closer to me: Spatial proximity amplifies interpersonal liking. *Personality and Social Psychology Bulletin*, 45(2), 300-309.
- Shrauger, J. S., & Schoeneman, T. J. (1979). Symbolic interactionist view of self-concept: Through the looking glass darkly. *Psychological Bulletin*, 86, 549–573. <https://doi.org/10.1037/0033-2909.86.3.549>
- Silk, J. S., Siegle, G. J., Lee, K. H., Nelson, E. E., Stroud, L. R., & Dahl, R. E. (2014). Increased neural response to peer rejection associated with adolescent depression and pubertal development. *Social cognitive and affective neuroscience*, 9(11), 1798-1807.
- Silva, C. S., Martins, A. C., & Calheiros, M. M. (2021). ‘Why do I think what I think I am?’: Mothers’ and fathers’ contributions to adolescents’ self-representations. *Self and Identity*, 20(8), 1015–1035. <https://doi.org/10.1080/15298868.2020.1792338>
- Silverstone, P. H., & Salsali, M. (2003). Low self-esteem and psychiatric patients: Part I – The relationship between low self-esteem and psychiatric diagnosis. *Annals of General Hospital Psychiatry*, 2(1), 2. <https://doi.org/10.1186/1475-2832-2-2>
- Slavich, G. M., O’Donovan, A., Epel, E. S., & Kemeny, M. E. (2010). Black sheep get the blues: A psychobiological model of social rejection and depression. *Neuroscience & Biobehavioral Reviews*, 35(1), 39–45. <https://doi.org/10.1016/j.neubiorev.2010.01.003>
- Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23, S208–S219. <https://doi.org/10.1016/j.neuroimage.2004.07.051>
- Somerville, L. H., & Casey, B. J. (2010). Developmental neurobiology of cognitive control and motivational systems. *Current opinion in neurobiology*, 20(2), 236-241.
- Somerville, L. H., Kelley, W. M., & Heatherton, T. F. (2010). Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. *Cerebral Cortex*, 20(12), 3005–3013.

- Sowislo, J. F., & Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychological Bulletin*, 139(1), 213–240. <https://doi.org/10.1037/a0028931>
- Stattin, H., & Skoog, T. (2016). Pubertal timing and its developmental significance for mental health and adjustment.
- Stendel, M., Guthrie, T. D., Guazzelli Williamson, V., & Chavez, R. (2023). Self-esteem modulates the brain-to-brain similarity of the representation of the self in the brains of others. *PsyArXiv*. <https://doi.org/10.31234/osf.io/tvdme>
- Strohming, N., & Nichols, S. (2014). The essential moral self. *Cognition*, 131(1), 159–171. <https://doi.org/10.1016/j.cognition.2013.12.005>
- Suh, E. M. (2007). Downsides of an Overly Context-Sensitive Self: Implications From the Culture and Subjective Well-Being Research. *Journal of Personality*, 75(6), 1321–1343. <https://doi.org/10.1111/j.1467-6494.2007.00477.x>
- Swallow, S. R., & Kuiper, N. A. (1988). Social comparison and negative self-evaluations: An application to depression. *Clinical Psychology Review*, 8(1), 55–76.
- Swann Jr, W. B., & Read, S. J. (1981). Self-verification processes: How we sustain our self-conceptions. *Journal of Experimental Social Psychology*, 17(4), 351–372.
- Tackman, A. M., Sbarra, D. A., Carey, A. L., Donnellan, M. B., Horn, A. B., Holtzman, N. S., ... & Mehl, M. R. (2019). Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of personality and social psychology*, 116(5), 817.
- Talaifar, S., Buhrmester, M. D., Ayduk, Ö., & Swann, W. B. (2021). Asymmetries in Mutual Understanding: People With Low Status, Power, and Self-Esteem Understand Better Than They Are Understood. *Perspectives on Psychological Science*, 16(2), 338–357. <https://doi.org/10.1177/1745691620958003>
- Tangney, J. P. (2002). Perfectionism and the self-conscious emotions: Shame, guilt, embarrassment, and pride. In G. L. Flett & P. L. Hewitt (Eds.), *Perfectionism: Theory, research, and treatment* (pp. 199–215). American Psychological Association.
- Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological bulletin*, 103(2), 193.
- Taylor, C. T., Bomyea, J., & Amir, N. (2011). Malleability of attentional bias for positive emotional information and anxiety vulnerability. *Emotion*, 11, 127–138. <https://doi.org/10.1037/a0021301>
- Taylor, T. L., & Montgomery, P. (2007). Can cognitive-behavioral therapy increase self-esteem among depressed adolescents? A systematic review. *Children and Youth Services Review*, 29, 823–839. <https://doi.org/10.1016/j.childyouth.2007.01.010>

- Tesser, A., Martin, L. L., & Cornell, D. P. (1996). On the substitutability of self-protective mechanisms.
- Tesser, A. (2003). Self-evaluation. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 275–290). The Guilford Press.
- Thomaes, S., Brummelman, E., & Sedikides, C. (2017). Why most children think well of themselves. *Child Development*, 88(6), 1873-1884.
- Thurston, M. D., Goldin, P., Heimberg, R., & Gross, J. J. (2017). Self-views in social anxiety disorder: The impact of CBT versus MBSR. *Journal of anxiety disorders*, 47, 83-90.
- Tibi-Elhanany, Y., & Shamay-Tsoory, S. G. (2011). Social cognition in social anxiety: First evidence for increased empathic abilities. *The Israel Journal of Psychiatry and Related Sciences*, 48(2), 98–106.
- Tice, D. M., & Wallace, H. M. (2003). The reflected self: Creating yourself as (you think) others see you. In *Handbook of self and identity* (pp. 91–105). The Guilford Press.
- Timbremont, B., & Braet, C. (2004). Cognitive vulnerability in remitted depressed children and adolescents. *Behaviour Research and Therapy*, 42(4), 423-437.
- Timbremont, B., Braet, C., Bosmans, G., & Van Vlierberghe, L. (2008). Cognitive biases in depressed and non-depressed referred youth. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 15(5), 329-339.
- Todd, A. R., & Tamir, D. I. (2024). Factors that amplify and attenuate egocentric mentalizing. *Nature Reviews Psychology*, 1-17.
- Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology*, 64, 231–255. <https://doi.org/10.1146/annurev-psych-113011-143812>
- Tone, E. B., & Tully, E. C. (2014). Empathy as a “risky strength”: A multilevel examination of empathy and risk for internalizing disorders. *Development and psychopathology*, 26(4pt2), 1547-1565.
- Tracy, A., Jopling, E., & LeMoult, J. (2021). The effect of self-referential processing on anxiety in response to naturalistic and laboratory stressors. *Cognition and Emotion*, 35(7), 1320-1333.
- Triandis, H. C. (1989). The self and social behavior in differing cultural contexts. *Psychological review*, 96(3), 506.
- Trzesniewski, K. H., Donnellan, M. B., Moffitt, T. E., Robins, R. W., Poulton, R., & Caspi, A. (2006). Low self-esteem during adolescence predicts poor health, criminal behavior, and limited economic prospects during adulthood. *Developmental Psychology*, 42, 381–390. <https://doi.org/10.1037/0012-1649.42.2.381>

- Tully, E. C., Ames, A. M., Garcia, S. E., & Donohue, M. R. (2016). Quadratic Associations Between Empathy and Depression as Moderated by Emotion Dysregulation. *The Journal of Psychology*, 150(1), 15–35. <https://doi.org/10.1080/00223980.2014.992382>
- Van de Groep, I. H., Bos, M. G. N., Jansen, L. M. C., Achterberg, M., Popma, A., & Crone, E. A. (2021). Overlapping and distinct neural correlates of self-evaluations and self-regulation from the perspective of self and others. *Neuropsychologia*, 161, 108000. <https://doi.org/10.1016/j.neuropsychologia.2021.108000>
- Van Der Molen, M., Poppelaars, E., Van Hartingsveldt, C., Harrewijn, A., Gunther Moor, B., & Westenberg, P. M. (2014). Fear of negative evaluation modulates electrocortical and behavioral responses when anticipating social evaluative feedback. *Frontiers in Human Neuroscience*, 7. <https://www.frontiersin.org/articles/10.3389/fnhum.2013.00936>
- van Schie, C. C., Chiu, C.-D., Rombouts, S. A. R. B., Heiser, W. J., & Elzinga, B. M. (2018). When compliments do not hit but critiques do: An fMRI study into self-esteem and self-knowledge in processing social feedback. *Social Cognitive and Affective Neuroscience*, 13(4), 404–417. <https://doi.org/10.1093/scan/nsy014>
- Van Tongeren, D. R., Davis, D. E., & Hook, J. N. (2014). Social benefits of humility: Initiating and maintaining romantic relationships. *The Journal of Positive Psychology*, 9(4), 313–321.
- Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological methods*, 16(1), 44.
- Vanderlind, W. M., Millgram, Y., Baskin-Sommers, A. R., Clark, M. S., & Joormann, J. (2020). Understanding positive emotion deficits in depression: From emotion preferences to emotion regulation. *Clinical psychology review*, 76, 101826.
- Vogt, K. S., & Norman, P. (2019). Is mentalization-based therapy effective in treating the symptoms of borderline personality disorder? A systematic review. *Psychology and Psychotherapy: Theory, Research and Practice*, 92(4), 441–464. <https://doi.org/10.1111/papt.12194>
- Vrieze, E., Pizzagalli, D. A., Demyttenaere, K., Hompes, T., Sienaert, P., de Boer, P., ... & Claes, S. (2013). Reduced reward learning predicts outcomes in major depressive disorder. *Biological psychiatry*, 73(7), 639–645.
- Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, 4(3), 274–290. <https://doi.org/10.1111/j.1745-6924.2009.01125.x>
- 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, J., Lüdtke, O., Robitzsch, A., Göllner, R., & Trautwein, U. (2018). Self-esteem development in the school context: The roles of intrapersonal and interpersonal social predictors. *Journal of Personality*, 86(3), 481–497. <https://doi.org/10.1111/jopy.12330>
- Walters, G. D. (2016). Reflected appraisals and self-view in delinquency development: An analysis of retrospective accounts from members of the Racine birth cohorts. *Journal of Criminal Justice*, 47, 100–107. <https://doi.org/10.1016/j.jcrimjus.2016.07.004>
- Washburn, D., Wilson, G., Roes, M., Rnic, K., & Harkness, K. L. (2016). Theory of mind in social anxiety disorder, depression, and comorbid conditions. *Journal of Anxiety Disorders*, 37, 71–77. <https://doi.org/10.1016/j.janxdis.2015.11.004>
- Watkins, E., & Teasdale, J. D. (2004). Adaptive and maladaptive self-focus in depression. *Journal of Affective Disorders*, 82(1), 1–8. <https://doi.org/10.1016/j.jad.2003.10.006>
- Watson, D., Clark, L. A., & Carey, G. (1988). Positive and negative affectivity and their relation to anxiety and depressive disorders. *Journal of abnormal psychology*, 97(3), 346.
- Wayment, H. A., & Taylor, S. E. (1995). Self-evaluation processes: Motives, information use, and self-esteem. *Journal of Personality*, 63(4), 729–757.
- Weary, G., & Edwards, J. (1994). Social cognition and clinical psychology: Anxiety, depression, and the processing of social information. In *Handbook of Social Cognition: Applications*. Psychology Press.
- Wegner, D. M. (2003). The mind's best trick: How we experience conscious will. *Trends in cognitive sciences*, 7(2), 65–69.
- Weinberg, A., Liu, H., Hajcak, G., & Shankman, S. A. (2015). Blunted neural response to rewards as a vulnerability factor for depression: Results from a family study. *Journal of abnormal psychology*, 124(4), 878.
- Weinstein, N. Y., Whitmore, L. B., & Mills, K. L. (2022). Individual Differences in Mentalizing Tendencies. *Collabra: Psychology*, 8(1), 37602. <https://doi.org/10.1525/collabra.37602>
- Weisbuch, M., Sinclair, S. A., Skorinko, J. L., & Eccleston, C. P. (2009). Self-esteem depends on the beholder: Effects of a subtle social value cue. *Journal of Experimental Social Psychology*, 45(1), 143–148. <https://doi.org/10.1016/j.jesp.2008.07.021>
- Wessels, N. M., Zimmermann, J., Biesanz, J. C., & Leising, D. (2020). Differential associations of knowing and liking with accuracy and positivity bias in person perception. *Journal of Personality and Social Psychology*, 118, 149–171. <https://doi.org/10.1037/pspp0000218>
- Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in neuroinformatics*, 7, 55610.

- Wild, L. G., Flisher, A. J., & Lombard, C. (2004). Suicidal ideation and attempts in adolescents: Associations with depression and six domains of self-esteem. *Journal of Adolescence*, 27(6), 611–624. <https://doi.org/10.1016/j.adolescence.2004.03.001>
- Will, G.-J., Moutoussis, M., Womack, P. M., Bullmore, E. T., Goodyer, I. M., Fonagy, P., Jones, P. B., Rutledge, R. B., & Dolan, R. J. (2020). Neurocomputational mechanisms underpinning aberrant social learning in young adults with low self-esteem. *Translational Psychiatry*, 10(1), Article 1. <https://doi.org/10.1038/s41398-020-0702-4>
- Will, G.-J., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *ELife*, 6, e28098. <https://doi.org/10.7554/eLife.28098>
- Williamson, V. G., & Mills, K. (2022). Mentalizing Strategies for Navigating the Social World in Adolescence. *PsyArXiv*. <https://doi.org/10.31234/osf.io/47ym2>
- Winer, E. S., & Salem, T. (2016). Reward devaluation: Dot-probe meta-analytic evidence of avoidance of positive information in depressed persons. *Psychological bulletin*, 142(1), 18.
- Winter, D., Steeb, L., Herbert, C., Sedikides, C., Schmahl, C., Bohus, M., & Lis, S. (2018). Lower self-positivity and its association with self-esteem in women with borderline personality disorder. *Behaviour Research and Therapy*, 109, 84–93. <https://doi.org/10.1016/j.brat.2018.07.008>
- Wong, Q. J., Gregory, B., & McLellan, L. F. (2016). A review of scales to measure social anxiety disorder in clinical and epidemiological studies. *Current psychiatry reports*, 18, 1-15.
- Wood, J. V., Heimpel, S. A., & Michela, J. L. (2003). Savoring versus dampening: Self-esteem differences in regulating positive affect. *Journal of Personality and Social Psychology*, 85(3), 566–580.
- Wright, J. C., Nadelhoffer, T., Perini, T., Langville, A., Echols, M., & Venezia, K. (2017). The psychological significance of humility. *The Journal of Positive Psychology*, 12(1), 3-12.
- Yaden, D. B., Haidt, J., Hood Jr, R. W., Vago, D. R., & Newberg, A. B. (2017). The varieties of self-transcendent experience. *Review of general psychology*, 21(2), 143-160.
- Yang, J., Dedovic, K., Guan, L., Chen, Y., & Qi, M. (2014). Self-esteem modulates dorsal medial prefrontal cortical response to self-positivity bias in implicit self-relevant processing. *Social Cognitive and Affective Neuroscience*, 9(11), 1814–1818. <https://doi.org/10.1093/scan/nst181>
- Yang, J., Xu, X., Chen, Y., Shi, Z., & Han, S. (2016). Trait self-esteem and neural activities related to self-evaluation and social feedback. *Scientific Reports*, 6, 20274. <https://doi.org/10.1038/srep20274>

- Yankouskaya, A., Bührle, R., Lugt, E., Stolte, M., & Sui, J. (2020). Intertwining personal and reward relevance: Evidence from the drift-diffusion model. *Psychological Research*, 84, 32-50.
- Yankouskaya, A., & Sui, J. (2021). Self-Positivity or Self-Negativity as a Function of the Medial Prefrontal Cortex. *Brain Sciences*, 11(2), Article 2. <https://doi.org/10.3390/brainsci11020264>
- Yoon, H.-J., Seo, E. H., Kim, J.-J., & Choo, I. H. (2019). Neural Correlates of Self-referential Processing and Their Clinical Implications in Social Anxiety Disorder. *Clinical Psychopharmacology and Neuroscience*, 17(1), 12–24. <https://doi.org/10.9758/cpn.2019.17.1.12>
- Yoon, L., Somerville, L. H., & Kim, H. (2018). Development of MPFC function mediates shifts in self-protective behavior provoked by social feedback. *Nature Communications*, 9(1), Article 1. <https://doi.org/10.1038/s41467-018-05553-2>
- Yue, C., Long, Y., Yang, Z., Xiao, Q., & Pan, W. (2021). The Influence of Actual Appraisals of Peers on the Self-Appraisals of Personality Traits for Chinese Late Adolescents: The Mediating Effect of Reflected Appraisals. *Frontiers in Psychology*, 12, 687482. <https://doi.org/10.3389/fpsyg.2021.687482>
- Yue, C., Yang, Y., He, W., Yue, T., & Pan, W. (2020). The Memory Effect of Reflected Self-Appraisals on Different Types of Others. *Frontiers in Psychology*, 11. <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.553585>
- Zabel, K., Goldbach, N., Fröhlich, U., Schneider, E., Görgülü, E., & Oertel, V. (2022). Enhancing self-esteem through cognitive behavioural group therapy—A randomized feasibility study of a transdiagnostic approach in psychiatric care. *Journal of Affective Disorders Reports*, 10, 100414. <https://doi.org/10.1016/j.jadr.2022.100414>
- Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological bulletin*, 146(2), 118.
- Zhang, D., Ao, X., Zheng, Z., Shen, J., Zhang, Y., & Gu, R. (2022). Modulating social feedback processing by deep TMS targeting the medial prefrontal cortex: Behavioral and electrophysiological manifestations. *NeuroImage*, 250, 118967. <https://doi.org/10.1016/j.neuroimage.2022.118967>
- Zhang, D., Shen, J., Bi, R., Zhang, Y., Zhou, F., Feng, C., & Gu, R. (2020). Differentiating the abnormalities of social and monetary reward processing associated with depressive symptoms. *Psychological Medicine*, 1–15. <https://doi.org/10.1017/S0033291720003967>
- Zhang, Q., Li, X., Wang, K., Zhou, X., Dong, Y., Zhang, L., Xie, W., Mu, J., Li, H., Zhu, C., & Yu, F. (2017). Dull to Social Acceptance Rather than Sensitivity to Social Ostracism in Interpersonal Interaction for Depression: Behavioral and Electrophysiological Evidence

from Cyberball Tasks. *Frontiers in Human Neuroscience*, 11.
<https://www.frontiersin.org/article/10.3389/fnhum.2017.00162>

Zheng, C., & Wu, Y. (2020). The more modest you are, the happier you are: The mediating roles of emotional intelligence and self-esteem. *Journal of Happiness Studies*, 21, 1603-1615.

Zhou, A., Li, S., Herbert, C., Xia, R., Xu, K., Xu, Q., Zhu, J., & Ren, D. (2013). Perspective taking modulates positivity bias in self-appraisals: Behavioral and event-related potential evidence. *Social Neuroscience*, 8(4), 326–333.
<https://doi.org/10.1080/17470919.2013.807873>

Zhou, J., Li, X., Tian, L., & Huebner, E. S. (2020). Longitudinal association between low self-esteem and depression in early adolescents: The role of rejection sensitivity and loneliness. *Psychology and Psychotherapy: Theory, Research and Practice*, 93(1), 54–71.
<https://doi.org/10.1111/papt.12207>.