1 Introduction and Overview

Widespread belief in the benefits of empathy is evident in public discourse and across diverse news and social media outlets around the world. Its potential as a remedy for an array of societal and relational problems such as aggression, intergroup conflict, and discrimination has clearly captured our collective imagination. Indeed, a recent Google search on the term "empathy" and its variants yielded approximately 64 million hits (exceeding "self-esteem" at 63 million). Further along these lines, promoting empathy is a key component of many interventions designed to improve interpersonal and intergroup relations. These interventions are diverse in many ways. They include, for example, multicultural education provided to college students, mobile phone applications such as the Random Acts of Kindness (RAKi) app (see www.rakigame.com), visits to school classrooms by a parent and baby as in the Roots of Empathy program, and role-playing exercises in which students or employees are arbitrarily divided into groups based on eye color and given firsthand experience with discrimination (Elliott, 2017). Despite their different approaches, these interventions share in common a goal of enhancing individuals' empathy for other people.

At first blush, such faith in the power of empathy would not seem misplaced. Substantial empirical research does indeed document that it can have numerous benefits. Moreover, of particular relevance to intergroup relations, there is at the same time evidence of an "empathy gap" across group lines whereby individuals feel more empathic toward members of their own group than toward outgroup members (e.g., Bruneau, Cikara, & Saxe, 2017; Cikara, Bruneau, & Saxe, 2011; Gutsell & Inzlicht, 2012). The conclusion seems obvious: reduce intergroup conflict by promoting empathy toward outgroup members.

However, a growing body of empirical and theoretical work has revealed that empathy can sometimes backfire in intergroup contexts, exerting negative rather than positive effects on individuals' attitudes and behavior toward others. For example, it can lead individuals to defensively derogate outgroup members in response to perceiving that outgroup members are critical of them (Vorauer & Sasaki, 2009), detract from intimacy-building behavior exhibited during back-and-forth intergroup interaction (Vorauer, Martens, & Sasaki, 2009), foster selfish behavior (Epley, Caruso, & Bazerman, 2006), and increase revenge seeking (Okimoto & Wenzel, 2011). This research highlights the need for a measured approach to promoting empathy in intergroup contexts that is sensitive to the conditions under which negative outcomes are likely. Also evident is a need to identify strategies for circumventing such negative effects so that empathy's potential to foster stronger social bonds across group boundaries can be more fully realized.

In the analysis that follows, I first review research documenting positive and negative effects of empathy and then consider implications for intervention in intergroup contexts. My analysis and recommendations emphasize the potential counterproductive influence of concerns with negative evaluation by outgroup members, building on previous research and theorizing regarding empathy and evaluative concerns (e.g., Vorauer, 2013) to also consider how such concerns might be mitigated. Notably, I focus on the implications of different types of interventions for warmth- and positivity-relevant outcomes such as willingness to interact, feelings of hostility, and treatment of outgroup members. However, I also consider power-relevant outcomes of minority groups and other kinds of outcomes that are often an important goal of interventions, such as enhanced appreciation of the outgroup's collective narrative. Thus, I generally focus on micro-level (intrapersonal) and meso-level (interpersonal) rather than macro-level (social structural) phenomena (see, e.g., Wright, Mazziotta, & Tropp, 2017). I conclude by integrating the evaluative concerns perspective with other analyses of empathy that have been advanced, maintaining an emphasis on the intrapsychic and social dynamics instantiated by empathy in intergroup interaction contexts. The evaluative concerns framework is notable for its heuristic value and the structure it provides: considering the connection of empathy to concerns with negative evaluation provides an overarching, theoretically driven set of recommendations for when and how to encourage empathy in intergroup contexts that is grounded in individuals' fundamental concerns with social acceptance and approval.

2 Research on the Effects of Empathy in Intergroup Contexts

2.1 Definition

What is empathy? Given that there is considerable variability in how empathy is construed by researchers and laypeople alike (see, e.g., Bloom, 2017; Cuff et al., 2016), there is no simple answer to this question. For the purpose of the present analysis, I adopt a working definition of empathy as an other-focused emotional response involving an orientation toward 'feeling for' another person. Notably, according to this definition, empathy does not require an accurate read of the target person's feelings or directly experiencing the presumed emotional state of the target person, which is sometimes referred to as "parallel empathy" (Stephan & Finlay, 1999). Further, my definition is consistent with the one advanced by Batson and colleagues, namely, "an other-oriented emotional response congruent with another's perceived welfare" (Batson, Polycarpou et al., 1997, p. 105), which involves feelings such as tenderness when the other is suffering. However, I incorporate the term "orientation" in my definition to explicitly include *efforts* to connect and identify with another person's feelings, in addition to a more spontaneous emotional response, as the former also involves a benevolent stance toward another person's feelings. Empathy manipulations in experimental research typically involve instructing individuals to imagine how another person is feeling, and measures typically involve asking individuals about the extent to which they feel sympathetic, compassionate, warm, and so on toward another person (as in Batson, Polycarpou et al., 1997).

Perspective-taking is a related construct – typically viewed as more cognitive in nature – that involves efforts to imagine another person's point of view by mentally stepping into his or her shoes and seeing the world through his or her eyes. As with empathy, perspective taking does not necessarily involve more accurate judgments of targets: accuracy is but one of several potential outcomes of perspective taking efforts, and indeed recent research suggests that actively trying to adopt another person's perspective generally does not result in more accurate judgments about that person (Eyal, Steffel, & Epley, 2018).

Although empathy and perspective taking are conceptually distinct and can have different effects (e.g., Galinsky, Maddux et al., 2008; Gilin et al., 2013; Vorauer & Quesnel, 2016), in practice they are closely intertwined: perspective taking can lead to empathy (e.g., Coke, Batson, & McDavis, 1978; Vescio, Sechrist, & Paolucci, 2003) and empathy can lead to perspective taking (Vorauer & Sasaki, 2009). The overlapping nature of these constructs is further illustrated by the fact that in the research literature, perspective taking instructions are sometimes involved in empathy manipulations (e.g., Galinsky, Maddux et al., 2008), and instructions to focus on another's feelings are sometimes included in perspective-taking manipulations (e.g., Davis et al.,1996; Vescio et al., 2003). Accordingly, for the sake of comprehensiveness, I draw on perspective taking as well as empathy research in this review. I will not dwell on the distinction except where it is relevant to the outcomes being considered and when I discuss different types of perspective taking toward the end of my analysis.

2.2 Positive Effects

Before delving into a review of circumstances in which empathy has been shown to have negative effects, it is important to acknowledge that a large research literature documents that it can often have positive effects. In particular, empathy has been clearly linked to helping behavior; ample evidence also indicates that it can foster a sense of a bond with others in the form of self-other merging, whereby self and other overlap in individuals' hearts and minds (for more thorough reviews, see Batson, Ahmad, & Lishner, 2009; Galinsky, Ku, & Wang, 2005; Hodges, Clark, & Myers, 2011; Vorauer, 2013).

Especially relevant for the current analysis, a large number of studies suggest that empathy can have positive implications for intergroup relations. For example, in a now-classic study, Batson, Polycarpou et al. (1997) found that inducing individuals to feel empathy for a member of a stigmatized group (e.g., a homeless man or a woman with AIDS) led them to report more positive attitudes towards the stigmatized group as a whole (see also, e.g., Broockman & Kalla, 2016; Finlay & Stephan, 2000; Shih, Stotzer, & Gutiérrez, 2013). Similar effects have been documented in the context of conflictual intergroup relations. In one study, Pliskin et al. (2014, Study 1) found that Jewish Israelis who were induced to feel empathy toward a West Bank Palestinian boy through reading that he had been diagnosed with cancer reported greater support for conciliatory policies toward Palestinians in general, although this effect was limited to those with a leftist political orientation. In a similar vein, a correlational study indicated that Jewish Israelis' empathy toward Palestinians was negatively correlated with support for aggression as part of the Israeli-Palestinian war, with this relationship being particularly strong for those with a leftist political orientation (Pliskin et al., 2014, Study 5). Other investigations have also found a negative association between Jewish Israelis' empathy for Palestinians and support for aggressive policies and actions during the war in Gaza (Rosler, Cohen-Cohen, & Halperin, 2017).

Work by Galinsky, Todd, and colleagues extends these findings to implicit intergroup attitudes and a range of information-processing outcomes related to reliance on stereotypes (e.g., Galinsky & Moskowitz, 2000; Todd et al., 2011). For example, Todd, Galinsky, and Bodenhausen (2012) found that perspective taking enhanced individuals' recall of an outgroup member's stereotype-inconsistent behaviors and led them to make more internal attributions for such behaviors; perspective taking also enhanced their pursuit of stereotype-inconsistent information. Other research has demonstrated that empathy can enhance the perceived injustice of discrimination toward minority group members (Dovidio et al., 2004). Although effects are not always positive (e.g., Lai et al., 2014; Mooijman & Stern, 2016), results like these make empathy attractive as a tool for intervention.

Notably, however, studies documenting positive effects in the intergroup domain have typically involved abstract (and ambiguous) "absentee" targets who are not physically present and who are instead represented by a photograph, transcript, or video clip. Beyond not being well positioned to pin down cause and effect, correlational studies examining attitudes toward the outgroup as a whole also involve abstract targets who are not physically present. Moreover, where behaviors rather than attitudes have been examined, there have typically been clearly delineated response options whose desirability centers on warmth and is unequivocal. For example, individuals may be asked to sign a petition or vote in favor of allocating resources to an agency that helps an outgroup in need (e.g., Batson et al. 2002; Bruneau et al., 2017), be given an opportunity to directly contribute money to an outgroup cause (Bruneau et al., 2017), or be asked how they would respond to a hypothetical direct request for help from an outgroup member (Sierksma, Thijs, & Verkuyten, 2015). Many real-world circumstances in which it might seem worthwhile to encourage empathy are messier. Moreover, although it is also important to consider consequences for the experiences and power of those on the receiving end of empathy, these outcomes are often neglected in research.

Thus, research documenting positive effects of empathy has generally involved a restricted set of conditions and outcomes that do not always seem to correspond well to the types of contexts – involving conflictual intergroup interaction – in which it may at first blush seem most needed and desirable as an intervention. Indeed, much of the evidence of backfiring effects comes from studies involving the potential for evaluation, complex behavioral response options, outcomes relevant to target experience and power, or some combination of these. Ultimately, considering the experimental contexts in which positive versus negative effects have been demonstrated enables predictions about real-world contexts where empathy is more versus less likely to be beneficial. Such an analysis also points to how the likelihood of negative effects might be minimized.

2.3 Negative Effects: Contributing Factors

Evidence for negative outcomes comes from research involving ethnic groups that occupy different positions of power in society as well as from research involving other types of group memberships such as those based on university affiliation or experimental groups created in the laboratory. The negative effects of adopting an empathic mind-set that have been documented are diverse. In terms of empathizers' reactions, they include activation of negative beliefs about how the outgroup views the ingroup, negative evaluations of the outgroup, revenge seeking, and engaging in selfish and unethical behavior. For targets, negative effects can involve a reduced psychological sense of power. Under what conditions are such negative effects most likely?

2.3.1 Potential for Negative Evaluation

Individuals in conflict with members of another group may well be in direct contact with outgroup members and readily identifiable to them. In such cases, there is clear potential for them to be evaluated by outgroup members, meaning that outgroup members are in a position to form impressions and make judgments about them personally. Even outside of such circumstances, explicit reference to intergroup judgment or topics that lead individuals to imagine interacting with outgroup members can also raise the specter of evaluation. Moreover, because individuals are generally more sensitive to the possibility of negative than positive evaluation (Leary & Downs, 1995) and have a basic appreciation of people's tendency to be more favorable toward ingroup than outgroup members (i.e., ingroup bias), in intergroup contexts the potential for evaluation typically translates in individuals' minds into alertness to the possibility of negative evaluation in particular.

Regardless of how it is instantiated, the potential for negative evaluation is likely to interfere with the positive effects of empathy and make negative effects more likely. Why might this be? Unlike "abstract" empathy applied to physically removed targets, empathy adopted toward an outgroup member who is in a position to evaluate them is apt to activate individuals' fundamental concerns with social evaluation and acceptance and bring such concerns to the foreground of their attention. Because of individuals' basic egocentrism (Zuckerman et al., 1983) and motivation to know and manage whether they are accepted or rejected by others (Leary & Downs, 1995), when they try to step into an outgroup member's shoes and empathize and imagine his or her feelings, they are likely to become focused on gauging his or her thoughts and feelings about *them*. For example, if their own group is relatively advantaged, they might imagine criticism attached to historical injustice and wrongs perpetrated against the outgroup by their own group or resentment attached to ongoing discrimination and inequality. If their own group is relatively disadvantaged, they might imagine being disrespected or dehumanized by the outgroup. Even if there is no particular power differential, many of these concerns could still apply. More generally, in connection with any type of group membership, individuals may consider stereotype-based expectations that the outgroup may have about them. Although such a focus on negative possibilities may seem incongruent with the prosocial orientation associated with empathy, it is highly congruent with research in social psychology underscoring individuals' fundamental preoccupation with monitoring their social standing with others – and particular attention to the possibility of negative evaluation.

Much of the evidence for the moderating role of concern with evaluation is indirect, resting on a comparison across studies that show positive effects of empathy (which generally do not involve potential for evaluation) and studies that show negative effects (which generally do involve potential for evaluation). However, one especially relevant experiment by Vorauer and Sasaki (2009) was designed to directly test the moderating role of the potential for evaluation by an outgroup member. In this experiment, White Canadian university students (that is, Canadian students with a European ethnic background) exchanged written information about themselves with an ostensible partner in the study who was depicted as either White or Indigenous Canadian. Thus, when their ostensible partner was Indigenous, participants were in the position of potentially being evaluated by an outgroup member, whereas this was not the case when their ostensible partner was White. The written information involved describing their personal qualities and answering questions from Aron et al.'s (1997) closeness-inducing procedure (e.g., "If you could change anything about the way that you were raised, what would it be? Why?"). Halfway through the written exchange, participants viewed a segment of a documentary depicting hardships endured by Indigenous Canadians in Northern Manitoba (*Wrapped in Plastic: Housing Manitoba First Nations*). The segment focused on the abysmal living conditions experienced by an Indigenous woman and her family in a northern Manitoba community. This aspect of the study thus involved presenting all participants with an outgroup member who was physically removed and in no position to identify or evaluate them. Following Batson, Polycarpou et al. (1997), participants in the objective condition were instructed to remain objective and detached while viewing the documentary segment, whereas those in the empathic condition were instructed to imagine the woman's feelings. A manipulation check confirmed that those in the empathy condition felt more empathy and liking for the woman in the documentary than did those in the objective condition.

Consistent with predictions, the results indicated that when participants empathized with the Indigenous woman in the documentary in the midst of a personal exchange with an Indigenous person (i.e., empathy in intergroup interaction), they activated negative meta-stereotypes about how White Canadians are viewed by Indigenous Canadians (e.g., *prejudiced, cruel, unfair,* *selfish*): responses to meta-stereotype–relevant words in a lexical decision-making task were quicker in this condition than they were when the empathy toward the Indigenous woman in the documentary was enacted in the midst of an interaction with a fellow White Canadian or when participants adopted an objective stance toward the Indigenous woman in the documentary. There were no such effects on stereotype-irrelevant words (e.g., *dishonest, pessimistic*) that were also included in the lexical decision-making task. In addition, stereotypes about the outgroup (e.g., *lazy, irresponsible*) were activated when those low in public collective self-esteem, who generally considered their ingroup to be viewed relatively unfavorably, were prompted to empathize with the Indigenous woman in the documentary. Further, although either intergroup contact or empathy toward the outgroup alone had prejudice-reducing implications, the combination – empathy with an outgroup member in the context of an intergroup interaction – did not. Empathizing with the Indigenous woman in the documentary in the context of intergroup interaction also reduced higher-prejudice individuals' desire for future interaction with their Indigenous partner in the study and led them to perceive that their partner was less interested in future interaction with them.

Other studies in which negative effects have been obtained in the context of intergroup relations have also involved the potential for evaluation. Consider, for example, an intervention carried out over a year in a conflict zone (Eastern Democratic Republic of Congo) in the form of a talk show that encouraged taking outgroup members' perspectives (broadcast in connection with a radio soap opera). This intervention, involving approximately fifteen ethnic groups, was found to have a range of negative effects including less tolerance of a disliked group and less willingness to help them by giving them salt, a valued commodity (Paluck, 2010).

A set of studies by Tarrant, Calitri, and Weston (2012) exploring perspective taking in the context of university and national (British versus German) group membership provides a particularly interesting case. These authors used a perspective taking task frequently used by Galinsky, Todd, and colleagues involving describing, as if they were the outgroup member, a day in the life of an outgroup member depicted in a photograph (see, e.g., Galinsky & Moskowitz, 2000). Tarrant et al., who explicitly told their participants that the research focused on intergroup evaluation, found that those high in ingroup identification evaluated outgroup members more negatively if they had been prompted to take the outgroup's perspective than if they had not. Because Galinsky, Todd, and colleagues typically find positive effects and typically present the manipulation and (often implicit) measures to participants as unrelated judgment and decision-making tasks, it is tempting to conclude that the salience of intergroup evaluation helps account for the divergent results obtained. It also seems plausible that those higher in ingroup identification would be more sensitive to the possibility of criticism from other groups. However, as Tarrant et al. do not present data on underlying process and suggest a different account, this analysis is speculative.

Further evidence for backfiring comes from contexts that involve conflict or competition and where the other party's potential negative evaluation thus looms large. In terms of conflict, Okimoto and Wenzel (2011) found that when individuals were instructed to empathize with the feelings of someone who had purposefully treated them negatively, they were more rather than less likely to seek revenge against the person. Although this possibility was not assessed, it seems likely that these results were due in part to increased focus on the other's apparently negative evaluation of them, that is, more energy that individuals devoted to thinking about how they were disliked, disrespected, or disregarded by the other person.

In terms of competition, Epley, Caruso, and Bazerman (2006) found that when individuals were encouraged to take the perspective of members of another group with which their own group was competing, they activated theories regarding others' likely selfish inclinations that in turn made them behave more selfishly: Epley et al.'s results, which were obtained with temporary groups created within an experimental context, indicated that considering a competitive outgroup's perspective led individuals to think about that group's likely negative inclinations toward them and to then respond in kind. For example, in one study taking the perspective of members of another group increased individuals' belief that members of that group would exaggerate their need when seeking to draw on a common resource and in turn increased individuals' own efforts to draw on the resource. Pierce et al., (2013) obtained conceptually parallel results with respect to unethical behavior.

Two final points are of note here. First, as mentioned previously, in the context of intergroup relations, salient potential for evaluation typically means salient potential for negative evaluation in particular. However, as articulated in greater detail later, even when individuals imagine a more positive potential evaluation, as when the relationship is cooperative or individuals have favorable intergroup attitudes, positive effects of perspective taking have failed to materialize (as in Pierce et al., 2013) and negative effects have sometimes been documented (as in Vorauer, Martens, & Sasaki, 2009). Nonetheless, this analysis focuses on concerns about negative evaluation because such concerns are most apt to characterize contexts in which effective interventions are sought and because these concerns are apt to more reliably have harmful consequences.

Second, concerns with negative evaluation can interfere with empathic responsiveness in two key ways. One possibility is that efforts to empathize lead to thoughts about negative evaluation that preclude the experience of empathy and associated reactions such as self-other merging in the first place. In essence, the process is hijacked: instead of empathic feelings, other reactions such as defensive derogation ensue. However, it is also possible for feelings of empathy to coincide with different kinds of negative reactions such as discomfort, guilt, and a desire to avoid outgroup members (see, e.g., Vorauer & Sasaki, 2009). Although these reactions may not involve antipathy, they can nonetheless be highly problematic, as they have implications for the inclusion of outgroup members across diverse social and employment contexts. It may be easier and less stressful for individuals to restrict their interactions to ingroup members and thereby avoid the negative evaluations they imagine wherever they have the power to do this. Further, discomfort and inhibition may well be interpreted by outgroup members as reflective of antipathy (see Devine, Evett, & Vasquez-Suson, 1996). Both efforts to empathize and the actual experience of empathic feelings may thus have negative consequences as a function of concerns with being seen in an unfavorable light by outgroup members.

2.3.2 Complex Behavioral Response Options and Ambiguity

Intergroup exchanges are often complex and characterized by considerable ambiguity. In particular, direct contact in the form of face-to-face or even computer-mediated exchanges typically provides a broad range of behavioral response options: individuals could be passively or actively aggressive, try to be helpful, decide to directly refer to intergroup relations and issues or avoid such issues altogether, and so on. Moreover, considerable ambiguity can surround the appropriate interpretation of behavior. It can be unclear, for example, whether a remark reflects hostile or defensive intentions. Even seemingly prosocial overtures such as providing help may come across to the recipients as controlling or patronizing instead of indicative of warm feelings or respect. Especially relevant to the present analysis, individuals may be uncertain about the signals their own behavior communicates to outgroup members. They may wonder, for example, whether making eye contact and asking questions will come across as attentive or aggressive, or whether being reserved will seem respectful or avoidant.