Title
Distinguishing Between Ideas and Experiences in Interpersonal Evaluation

Author
Wang, Yilin

Publication Date
2021

Peer reviewed|Thesis/dissertation

Distinguishing Between Ideas and Experiences in Interpersonal Evaluation

By

YILIN ANDRE WANG

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____

Alison Ledgerwood, Chair

_____

Paul W. Eastwick

_____

Jeffrey W. Sherman

Committee in Charge

2021

Acknowledgements

I came across this exercise when I studied choreography in college. The exercise was simple: write down every single sound you can hear. The first time I tried it, I was sitting by a windowsill in my parents' apartment in early February. Night had fallen, but the birds had not yet rested. Their intermittent songs weaved with the swoosh of cars on a nearby boulevard, the static of the streetlamps, the hum of the fridge in the kitchen, the gentle tick of the alarm clock by the bed, and the low thuds from the neighbors upstairs. As I wrote down more and more sounds, I was struck by the richness of the sonic landscape around me. It was like I was learning to hear for the first time.

Writing these acknowledgements for my dissertation is, in many ways, the same exercise. It gives me a moment to intentionally take stock of the many voices that have surrounded and guided me, the voices that accompanied and shaped my academic path, the voices that I turn to for insight and comfort. I am awash with gratitude as I reflect on these voices and what they mean to me; here, I humbly attempt to respond with my own.

To my graduate advisor Alison Ledgerwood, I can't find enough words to express the many ways that you have made an impact on me. You inspire me to treat people with open heart, to work with integrity, to share my best, and to celebrate the collective good that I am a part of. I am grateful that you see me as I am, and that you helped me grow so much throughout my time in grad school. I came to grad school to become a scientist; I graduate having learned from you not only how to do good science, but how to do good by scientists. I hope to pay it forward for my future students and be their champion as you have been mine.

To Paul Eastwick, Andy Todd, and Mijke Rhemtulla, thank you for your sterling insights, support, enthusiasm, and kindness. It has been a pleasure working with and learning from you; I

am grateful for your presence, and I hope we will continue our connections for many years to come. To Jeff Sherman, thank you for introducing me to social cognition and for pushing me to think deeply over the years. To Cindy Pickett, thank you for seeing my potential. To Simine Vazire, Chris Hopwood, Wiebke Bleidorn, Karen Bales, and Kristin Lagattuta, thank you for inspiring me with your scholarship and for your wise advice. To Saaid Mendoza, Catherine Sanderson, and Lisa Raskin, thank you for introducing me to the wonders of psychology.

I would not have survived graduate school without the friendship that I have made with my peers. Jehan, Shannon, Mark, and Amber, thank you for showing me the ropes in the lab, for your intellect, and for your support. Aline, Chris, and Azra, I am so glad our time at Davis overlapped, and I wish the very best for your future endeavors. To the Ledgerwood lab managers and research assistants, thank you for being part of the research team—I wish I could list all your names here. To my cohort mates, especially Hannah, Gent, and Heather R., I am so glad we made it! Thank you, Ted, Alexia, Lynea, Leigh, Heather M., Angela, Sam, Jessie, Katie, Ryan, Micheal, Jesse, and so many other folks I met in grad school, for your companionship, warmth, and presence. Thank you, Alice and Alex, for being my superheroes. I am also grateful to the many brilliant people I have met in this field: Being able to connect with you and engage with your minds is a beautiful thing I wish to never end. These pages do not allow me to do justice to the many communities that I have had the privilege of being a part of, but: Thank you to those who share with me happiness, peace, understanding, courage, redemption, growth, and the artistry of living.

To my parents: Thank you for making me who I am.

To Forrest: Thank you for being the light.

Abstract

From intelligence in romantic partners to empathy in political leaders, people can readily think about what they like in other people. These abstract ideas about liking can be useful in many situations, but do they always align with people's concrete experiences of liking? In this dissertation, I distinguish between ideas and experiences in interpersonal evaluation, and I argue that this distinction is useful because they can predict different kinds of decisions and help explain social psychological phenomena that appear paradoxical. In Chapter 1, I situate ideas versus experiences in a broader theoretical framework on attributes, and I conducted a series of studies showing that (1) ideas about liking for personal attributes can be affected by incidental features of the context, and that (2) ideas about liking versus experiences of liking predict different outcomes in the context of romantic attraction. In Chapter 2, I zero in on empathy and consider why people like the idea of empathy but not always those who show it. Across seven experiments, I found evidence that empathy has evaluative consequences for the empathizer beyond the empathic dyad. Findings from these experiments suggest that although people are often encouraged to empathize with disliked others, they are not always favored for doing so. In Chapter 3, I present a discussion and tutorial on a statistical technique central to my empirical work on ideas versus experiences, structural equation modeling (SEM). Specifically, I conducted a simulation study showing key factors that influence statistical power to detect true effects in SEM, and I introduce a free online app that helps researchers conduct power analysis for SEM. Taken together, these three chapters offer theoretical, empirical, and methodological advances for the study of ideas versus experiences in interpersonal evaluations. Unpacking the distinct roles that ideas versus experiences play can help us understand the apparent disconnect between what people *think* they like and what *drives* their liking across many domains.

Table of Contents

Chapter 1

Experiences of Liking and Ideas about Liking: Distinguishing Functional and Summarized Preferences for Partner Attributes

Abstract

People have ideas about the attributes (i.e., traits or characteristics that vary along a dimension) that they like in others (e.g., "I like intelligence in a romantic partner"), and these attitudes are called *summarized attribute preferences* (Ledgerwood et al., 2018). Do summarized attribute preferences capture the extent to which people actually experience liking for an attribute (called a *functional attribute preference*), and how do these two distinct forms of attribute preferences predict situation selection? Across four studies, we showed participants a series of photographs of faces and assessed both their experienced liking for an attribute (their functional attribute preference) as well as their inference about how much they liked the attribute in the abstract (their summarized attribute preference). Our results suggest that (1) summarized attribute preferences may be grounded—albeit weakly—in functional attribute preferences, and that (2) summarized attribute preferences can also be affected by incidental aspects of the context in which people learn about them. Furthermore, we observed a double dissociation in the predictive validity of summarized and functional attribute preferences: Whereas summarized attribute preferences predicted situation selection at a distance (e.g., whether to join a new dating website based on a description of it), functional attribute preferences predicted situation selection with sampling (e.g., whether to join a new dating website after sampling it). We discuss theoretical and methodological implications for the interdisciplinary science of human evaluation.

Keywords: stated preferences, abstraction, covariation detection, situation selection, attraction

Introduction

Preferences for attributes are central to the way that people think about and experience the world. A person might profess their love of spiciness in food or their appreciation for intelligence in a romantic partner; someone might be drawn to an area of the country where residents are more liberal or more conservative; the brightness of an apartment might drive a person's interest in signing a lease. Perhaps unsurprisingly, multiple literatures have studied such preferences for *attributes*—that is, qualities that vary along a continuum (Anderson, 1971; Borgia, 1995; Buss, 1989; Eastwick et al., 2014; Fishbein & Ajzen, 1975; Lawless & Heymann, 2010).

Notably, these interdisciplinary literatures contain two very different ways of conceptualizing and measuring attribute preferences (Ledgerwood et al., 2018). One common approach to understanding attribute preferences has focused on how strongly a given attribute is associated with liking. This association is called a *functional attribute preference*, and it is characterized as a (within-person) valenced response to increasing levels of an attribute in a series of targets (e.g., the extent to which intelligence in a series of romantic partners predicts a person's liking for each partner; Wood & Brumbaugh, 2009). Functional preferences are the primary target of investigation by researchers who study attribute preferences in nonhuman animals (e.g., birds; Borgia, 1995; Moller, 1988; Patricelli et al., 2002); for example, researchers interested in understanding mate preferences in satin bowerbirds assess female birds' functional preference for vocal mimicry ability in a mate by measuring the strength of the association between (a) the accuracy and size of male birds' vocal mimicry repertoires and (b) the males' courtship success (Coleman et al., 2007). Some human literatures emphasize functional

preferences, too (e.g., consumer preferences; Delgado & Guinard, 2011; Lawless & Heymann, 2010; organizational preferences; Heilman & Saruwatari, 1979; Turban & Keon, 1993).

Importantly, because humans can also abstract and articulate their likes and dislikes, a second approach to understanding attribute preferences is popular when studying humans (Anderson, 1968; Buss, 1989; Fletcher et al., 1999). This approach focuses on people's overall, summary evaluations of a given attribute in the abstract: A *summarized attribute preference* is a valenced response to an attribute as a concept (e.g., "I like intelligence in a romantic partner" or a negative gut reaction to the idea of ambitiousness in a leader). Summarized preferences are the primary target of investigation by researchers who study human mate preferences (e.g., Buss, 1989; Christensen, 1947; Fletcher et al., 1999, 2000; Hill, 1945), as well as preferences for attributes of friends and family members (Apostolou, 2007; Goodwin & Tang, 1991; see also Huang et al., 2020). For example, researchers in the fields of family studies, close relationships, and evolutionary psychology assess people's preferences for various attributes in a romantic partner by asking participants to rate how much they like or desire each attribute on a rating scale (e.g., a participant might rate the desirability of *attractiveness* or *intelligence* in a mate as a "7" on a 9-point scale ranging from 1 = *not at all desirable* to 9 = *extremely desirable*).

Because functional preferences and summarized preferences are largely studied in separate research traditions, much remains unknown about their relations. Researchers assessing summarized preferences often seem to use them as proxies for functional preferences (e.g., Gerlach et al., 2019), and it seems plausible that people's ideas about their likes and dislikes draw from their experienced evaluations in the moment, at least to some extent (Ledgerwood et al., 2018). However, because summarized preferences reflect abstract, summarized ideas about

liking—perhaps a uniquely human evaluative phenomenon—they may behave differently from functional preferences in terms of their antecedents and consequences.

Attitudes towards Objects and Attributes

Our examination of functional and summarized attribute preferences can also be situated in the context of the social psychological literature on attitudes. Common definitions of *attitude* in this literature include "a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (Eagly & Chaiken, 1993, p. 1) and "associations between a given object and a given summary evaluation of the object" (Fazio, 2007, p. 608). The terms "entity" and "object" were surely intended to be broad enough to capture attributes such as spiciness, intelligence, and other traits or characteristics that refer to dimensional qualities. But in practice, most research in the attitude literature has focused on the antecedents and consequences of people's evaluations of people, places, and things (see Ledgerwood et al., 2018, for a review and in-depth discussion).

Scholars in the social-psychological attitudes tradition might begin with the reasonable assumption that our knowledge base on attitudes towards objects would generalize to attitudes towards attributes. In other words, the vast literatures on persuasion, attitude structure, attitude strength, direct vs. indirect measurement, and so forth—literatures that have been honed by studying attitudes toward objects ranging from social groups to comprehensive exams to squirrels (e.g., Roskos-Ewoldsen & Fazio, 1992; Chaiken & Ledgerwood, 2012)—should apply to attitudes towards attributes like spiciness, intelligence, and brightness. Although this is a reasonable starting assumption, it is worth differentiating attitudes towards objects and attributes for at least two reasons.

First, there is a central theoretical perspective in this literature that posits distinct roles for attitudes towards attributes and attitudes towards objects. In classic expectancy-value models of attitude formation and change (Anderson, 1971; Fishbein & Ajzen, 1975; Lampel & Anderson, 1968), a person's attitude toward an attribute is positioned as an antecedent of (and is therefore distinct from) their attitude towards an object. In these models, an attitude toward an object depends on (a) the extent to which various attributes characterize the object (i.e., expectancy) and (b) the person's evaluations of these attributes (i.e., value, classically measured as a summarized attribute preferences). For example, a person's attitude toward an apartment might depend on (a) the extent to which they believe the apartment is bright, spacious, and centrally located and (b) the extent to which they positively evaluate the attributes of brightness, spaciousness, and centrality of location in an apartment. Notably, whereas extensive research has investigated the precursors of expectancy beliefs (e.g., Fishbein & Ajzen, 1975; Kaplan, 1973), very little research has investigated the precursors of attribute preferences: Attempts to examine causes of attitudes towards attributes are uncommon, perhaps because these attitudes initially proved resistant to manipulation attempts (Eagly & Chaiken, 1993; Eastwick et al., 2019; Lutz, 1975). In other words, we do not know much about what changes people's attribute preferences.

Second, objects and attributes are conceptually distinct because attributes—but not objects—contain their own natural contrast (Ledgerwood et al., 2018). Because attributes are dimensions, a given attribute contains a higher versus lower level contrast within itself (e.g., higher vs. lower levels of intelligence in a partner or spiciness in a food). Of course, objects can be contrasted with one another at a researcher's discretion (e.g., Coke vs. Pepsi or Coke vs. Sprite), but a given object does not typically have a single natural contrast by definition in the way that attributes do. The existence of a natural contrast for attitudes towards attributes presents

an additional complexity when people form novel attitudes toward attributes. Consider that the process of forming an attitude towards an object involves the weighing of positive and negative past experiences with the object (Fazio et al., 2015). Forming an attitude toward an attribute would further be informed by whether a person has (positive and negative) past experiences at *different levels of the attribute* across an array of objects. This dose-response association between the attribute and evaluative responses across objects is what we call a functional attribute preference, and it has no necessary logical parallel within the process of forming an attitude towards an object itself (Ledgerwood et al., 2018).[1]

Functional and Summarized Attribute Preferences

As discussed above, different research streams have tended to focus solely on either functional or summarized attribute preferences. Indeed, even Fishbein and Azjen (1975) focused solely on summarized preferences when assessing value, without considering the alternative possibility of using functional preferences to capture value. Furthermore, questions about summarized preferences only arise when studying humans, because humans, unlike other animals, readily exhibit summarized as well as functional preferences. For example, birds may exhibit a functional preference for vocal mimicry (i.e., vocal mimicry ability is positively correlated with mating interest), but unlike humans, they do not seem to form abstract ideas about the extent to which they like this quality in a mate. As a result, little empirical work has

[1] Critically, the difference between functional and summarized preferences is not merely a measurement distinction (see Ledgerwood et al., 2018, for a full discussion). Just like attitudes towards objects, both types of attribute preferences can be assessed in more direct or indirect ways. For example, one can measure summarized preferences for the attribute *intelligent* using a self-reported rating scale (i.e., a more direct measure) or using the relative reaction time to positive versus negative words after being primed with the word *intelligent* (i.e., a more indirect measure). Similarly, in a measure of functional preferences, both the intelligence of the targets and participants' liking for those targets can be assessed directly (i.e., rating scales) or indirectly (i.e., reaction times). Therefore, the distinction between summarized and functional preferences is not about direct versus indirect measurement but about whether participants are evaluating the attribute as a concept in and of itself (summarized) or are experiencing their liking for the attribute as instantiated in a set of targets (functional).

assessed both functional and summarized preferences, and the two have only recently been synthesized theoretically (Ledgerwood et al., 2018; see Figure 1.1). Such a synthesis prompts new questions about summarized preferences as a potentially uniquely human oddity: Where do humans' abstract ideas about their preferences come from, and what do these ideas predict? The current set of studies seeks to address these questions in the context of human mate preferences.

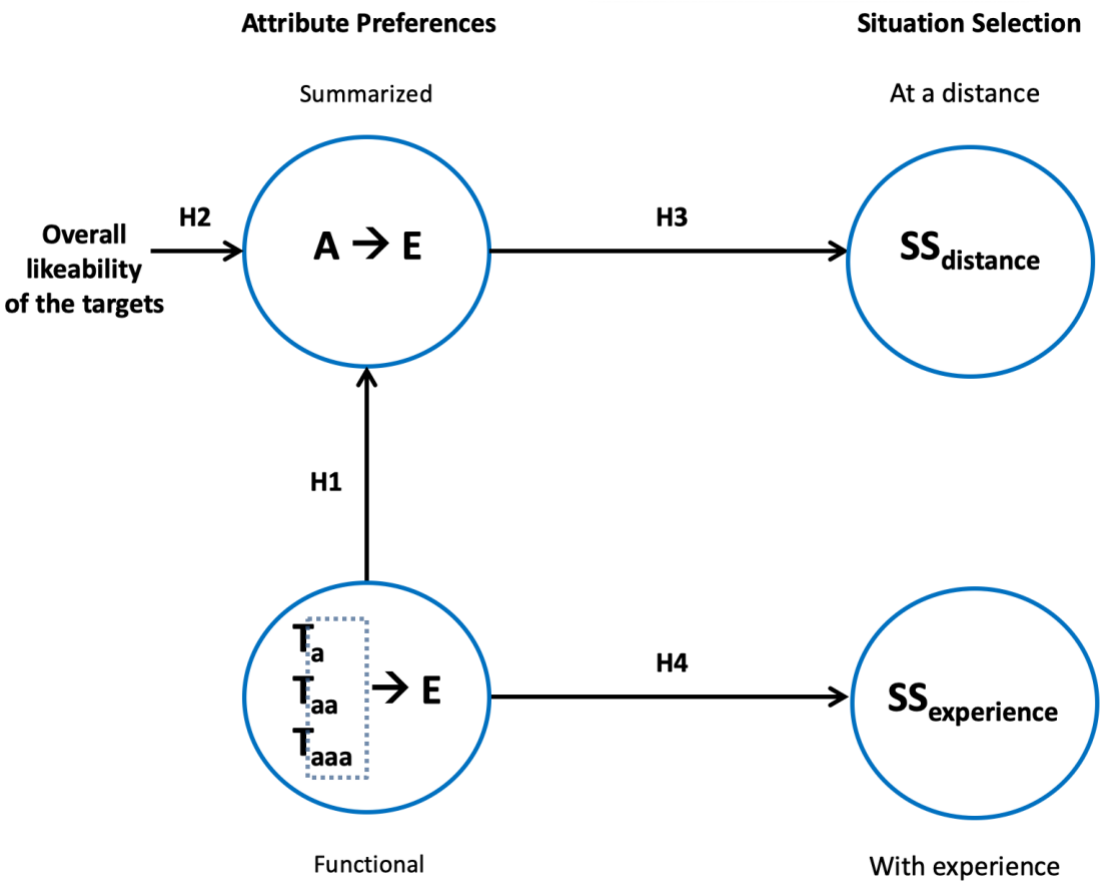

Figure 1.1. Theoretical framework of attribute preferences and the four hypotheses (H1–H4) tested in the current research, adapted from Ledgerwood et al. (2018, Models 1 and 3). A = attribute (as a concept; e.g., "intelligence in a romantic partner"); a = attribute (as exhibited by a target; e.g., a particular partner's level of intelligence); T = target person; E = evaluation; SS = situation selection. Summarized preferences reflect evaluations of the attribute as a concept; functional preferences reflect evaluations of increasing levels of the attribute in a series of targets.

From Functional to Summarized Preferences

According to our theoretical framework (see Figure 1.1), summarized preferences should be rooted in functional preferences to some extent—that is, logically, people's abstract ideas about the extent to which they like an attribute should be based on their experiences of liking for targets that vary along that attribute dimension. Indeed, many research literatures assume that they are linked (see Ledgerwood et al., 2018, for a review), and there is at least some empirical evidence for this assumption. In the only published experimental tests of this idea to date, Eastwick et al. (2019) manipulated functional preferences for an attribute and found that participants' summarized preferences changed in response to the manipulation. In nonexperimental studies with participants evaluating photographs of potential partners, functional and summarized preferences exhibit positive, small-to-moderate correlations in the handful of studies that have measured both types of preferences (Brumbaugh & Wood, 2013; Caruso et al., 2009; DeBruine et al., 2006; Eastwick & Smith, 2018; Wood & Brumbaugh, 2009). For example, in the largest of these studies, estimates of the correlations between functional preferences and summarized preferences across various traits ranged from $r = .02$ to $r = .38$ (Brumbaugh & Wood, 2013). Informed by these studies, we predicted that:

> H1: Summarized preferences for an attribute in a romantic partner will correlate with functional preferences for that attribute.

Critically, our decision to separate functional and summarized preferences and treat them as two separate constructs implies that they will be far from isomorphic (cf. Gerlach et al., 2019). In fact, there are reasons to expect that functional and summarized preferences might correlate quite weakly. For example, in research inspired by Brunswik's (1952) lens model, people's beliefs about the validity of cues in predicting outcomes (e.g., what behaviors people think indicate lying) can markedly diverge from the inferences they actually draw from those cues

(e.g., what behaviors they rely on for making judgments about lying; Hartwig & Bond, 2011). This divergence is consistent with research on the limitation of self-knowledge and insights into one's own decision-making process (e.g., Fiske & Taylor, 2008; Neisser, 1967; Nisbett & Wilson, 1977). Even though a person's functional preference (i.e., *actual* experienced liking for an attribute) would be a highly relevant piece of information for generating a summarized preference judgment (i.e., *beliefs* about liking for an attribute), people may not be particularly rational or accurate, and precise estimates of this association in different contexts remain rare.

Unique Antecedents of Summarized Preferences

If people do not perfectly abstract a summarized preference from their experienced functional preferences, they may draw on other sources of information when forming a summarized preference. To the extent that people learn about their summarized preferences from their past experiences, this learning process may be similar to the process of inferring abstract associations from case-by-case experiences in covariation detection tasks (sometimes called "contingency learning tasks;" Allan & Jenkins, 1980; Jenkins & Ward, 1965; Mandel & Lehman 1998; see also Perales et al., 2005; Fiedler et al., 2009). Drawing from this rich literature allows us to identify extraneous contextual inputs that may influence summarized preferences. In other words, the antecedents of a summarized preference may include not just a person's experienced functional preference but also incidental aspects of the context in which they are forming the new summarized preference.

In a typical covariation detection task, participants encounter a series of trials in which cues and outcomes vary and then make an abstract inference about the nature of the association between a cue and an outcome. For example, participants might encounter a series of trials in which a chemical is present or absent (cue) and bacteria survive or not (outcome; Allan et al.,

2005). Participants would then make an overall judgment about the relation between the chemical's presence and bacterial survival. In such studies, participants' abstract judgments are regularly influenced by features of the learning context that are orthogonal to the actual experienced association between cue and outcome. One critical contextual influence is the probability or "density" of the outcome itself (i.e., whether the outcome is encountered more or less frequently in the series of trials). For example, when the probability of bacterial survival is generally high (vs. low), participants tend to infer a stronger relation between the chemical and bacterial survival, even though the actual experienced association between cue and outcome is identical across conditions. This contextual effect of outcome probability on abstract contingency inferences has been dubbed the *outcome density bias* (e.g., Blanco, 2017; Blanco et al., 2013; Matute et al., 2015; Vadillo et al., 2013).

The process of translating a functional to summarized preference is likely similar to the mental abstraction process that participants use in a typical covariation detection paradigm (see also Eastwick et al., 2019). In both cases, people experience an association (between cue and outcome or between attribute and target evaluation) and then make an abstract judgment about the strength of that association. For example, a person might experience greater romantic liking for potential partners who are higher in intelligence, and then make an abstract judgment about how much they like intelligence in a romantic partner. Thus, when inferring a new summarized preference, people may be influenced by the same contextual factors that influence covariation detection judgments, such as the density of the outcome—in this case, the average positivity of the evaluations people are experiencing. In other words, when targets are more (vs. less) likeable on average, people experience more positive evaluations and thus may infer stronger

summarized preferences for an attribute even when functional preferences (i.e., the actual association of levels of that attribute with liking) are constant.

H2: Summarized preferences for a novel attribute will be more positive in a learning context with high versus low likeability targets.

Although this prediction about a contextual input for summarized preferences seems sensible, it is worth considering that forming summarized preferences is different from detecting covariation in a typical paradigm in important ways. First, in a typical covariation detection paradigm, cues and outcomes are binary: They are either present or absent (for exceptions that used continuous cues/outcomes, see Chow et al., 2019; Marsh & Ahn, 2009). In contrast, in the domain of preferences, both traits and liking typically exist on a continuum. For example, a potential partner's level of confidence can continuously range from very low to very high, and a person's evaluation of the partner could also continuously range from strongly negative to strongly positive. Second, in typical covariation detection paradigms, the actual association between cue and outcome is solely determined by the experimenter: Participants' experiences are tightly controlled to be identical. In contrast, people naturally experience their own evaluative responses in the real world, and these responses vary from person to person. Therefore, it is unclear whether the causal influence of outcome density or other contextual factors on abstract judgments would emerge in realistic, complicated contexts where people are learning about their own summarized preferences. Because of these differences, it is possible that the effect of outcome density observed in covariation detection tasks—which typically use binary variables and tightly controlled cue-outcome associations—will not appear when people make inferences about their summarized preferences.

Unique Consequences of Summarized Preferences

If the antecedents of summarized preferences include incidental contextual inputs like the average likeability of a set of targets, researchers may wonder if functional preferences reflect people's "real" attribute preferences. Are summarized preferences simply crude and noisy proxies for functional preferences? Indeed, this argument has been levied against summarized preferences in past research (Eastwick & Finkel, 2008; Brumbaugh & Wood, 2013). On the one hand, such an argument is supported by the fact that functional preferences represent people's experienced evaluations of attributes; they capture the rich and complex information manifested across various encountered objects in reality. In contrast, summarized preferences require that people simplify the rich, complex information represented in functional preferences into a single, overall summary judgment. Arguably, researchers studying human mating moved from measuring functional preferences (used in the non-human mating literature; e.g., Thornhill, 1983) to measuring summarized preferences (used in almost all studies of human mate preferences; e.g., Fletcher et al., 1999) because directly asking people about their summary judgments is a quick-and-easy measurement option when studying humans. But if summarized preferences tend to capture incidental aspects of the learning context, one could argue that researchers should always measure functional preferences unless it is too onerous to do so.

On the other hand, this view of summarized preferences as simply a poor and potentially contaminated measure of functional preferences might be overly simplistic. Considerable research suggests that abstract representations guide decision-making at a distance (Gilead et al., 2020; Trope & Liberman, 2010), and recent theoretical work suggests that summarized preferences are relatively abstract (Ledgerwood et al., 2018; Ledgerwood et al., 2020). Drawing on these ideas (which we discuss in more detail before Study 2), we posit that one purpose of summarized preferences that distinguishes them from functional preferences is that summarized

preferences enable humans to select into situations at a distance, without having to directly experience those situations. We will therefore test whether summarized preferences predict how people respond to situations they have not yet directly experienced (e.g., situations learned about only through verbal communication with others).

H3: Summarized preferences will predict situation selection at a distance (i.e., deciding on situations before directly encountering them).

The Current Research

In the current research, we distinguished between functional and summarized preferences for partner attributes. We situated our studies in the context of mate preferences—an area in which attribute preferences have been extensively studied—because mate selection exemplifies a real-life, complex process in which people develop and frequently express summarized preferences. We developed paradigms that enabled us to examine both the formation of summarized preferences (Studies 1–2), their relation to functional preferences (Studies 1–4), and the downstream consequences that summarized and functional preferences predict (Studies 2–4). In Studies 1–3, participants learned about their preferences for a novel attribute displayed in a series of preferred-sex target faces. In Study 4, we measured participants' existing summarized preferences for familiar attributes.

Across these studies, we tested our three hypotheses. First, in all studies, we tested the correlation between functional and summarized preferences (H1). When people are asked to evaluate photographs of potential mates, large-scale prior studies focusing on existing mate preferences for familiar attributes generally find small-to-moderate functional–summarized preference associations (e.g., Brumbaugh & Wood, 2013; Eastwick & Smith, 2018; Wood & Brumbaugh, 2009). Studies 1–3 allowed us to examine the strength of this association when

participants learn about a novel attribute depicted in the faces of preferred-sex romantic partners (see also Study S1 in the supplemental materials), whereas Study 4 allowed us to examine it for the familiar attributes *intelligence* and *confidence*.

Second, drawing from the covariation detection literature, we examined the possibility that outcome density—in this case, the average level of liking experienced toward a set of preferred-sex targets—could be a contextual input for summarized preference formation without affecting functional preferences. We predicted that experimentally manipulating a set of target faces to be more (vs. less) likeable would lead participants to infer stronger summarized preferences for a novel attribute, even if functional preferences remained the same (H2; Studies 1 and 2; see also Study S1). In other words, participants might (mistakenly) infer that they like an attribute more when they happen to learn about their preference in a context that elicits more (vs. less) liking for the targets, even if the average association between the attribute and liking (i.e., participants' averaged functional preference) is the same across conditions.

Third, we examined whether summarized preferences might predict decisions about situations that people learn about through socially acquired knowledge, before personally experiencing the situation directly. That is, when people learn about a novel situation from others (as when people read a description of a dating website that features partners high on intelligence or athleticism), their summarized preferences might predict the situation they select. Therefore, we hypothesized that summarized preferences would predict situation selection when participants encounter a description of a novel situation involving an opportunity to date romantic partners high on particular attributes (H3). We tested this hypothesis for both a novel attribute preference in a tightly controlled learning context (Studies 2–3) and for existing attribute preferences in a more realistic online dating context (Study 4).

In the supplemental materials, we report additional data on H1 and H2 (Study S1), a study that validated the measures used to assess romantic interest (Study S2), and a pilot study that collected norming data on the stimuli used in Study 4 (Study S3). These studies are ancillary to the main set of studies; we refer to them below when relevant to the main studies.

Following recent calls to constrain researcher degrees of freedom using analysis plans (Nosek et al., 2018; Ledgerwood, 2018), we set and recorded ahead of time (and for Studies 3 and 4, publicly preregistered) our analysis plans, including power analyses, target sample size, inclusion and exclusion criteria, and planned data analyses. Additional analyses that were not planned ahead of time were reported as such below.

Study 1

We began by designing a paradigm that would allow us to examine how people initially form summarized preferences. We created a context in which participants learned about an ostensibly novel facial characteristic named "Reditry." In fact, Reditry was babyfacedness; we gave this visible attribute a novel, unfamiliar name to bypass any pre-existing semantic associations that participants might have with the term babyfacedness. In this task, participants saw a series of real faces from the Chicago Face Database (CFD; Ma et al., 2015). We told participants how much Reditry each face had and asked them to report their liking for each face. After participants experienced their liking for the entire series of faces with varying levels of Reditry, participants reported their overall, summarized preference for Reditry.

This paradigm allowed us to examine our two research questions about summarized preference formation. We tested whether participants' summarized preferences for Reditry would be informed by (1) their functional preferences for Reditry (i.e., the association between

each face's level of Reditry and the participant's liking for that face; H1) and (2) the average likeability of the targets encountered in the learning context (H2).

Method

Participants and power. One hundred and seven participants completed the study online through Amazon's Mturk platform. They were randomly assigned to one of two between-subjects conditions (low average likeability vs. high average likeability). We decided *a priori* to target a cell size of 50 participants per cell based on our lab's standard practice for minimum cell size when we are not sure what effect size to expect in a new line of research (the total number of completed surveys in Qualtrics ended up being slightly higher).

We decided to use female faces as stimuli in our first study, for simplicity. Because the study measured participants' romantic liking for the faces, we limited participants to those who reported being primarily attracted to women and who were 18–35 years old to match the apparent age range of our stimuli. We set and recorded the following *a priori* exclusion criteria: We would exclude participants who (1) gave an identical rating to all faces presented for measurement of functional preferences, and/or (2) provided a nonsensical response to a Winograd-like schema designed to filter out bots or inattentive participants. The numbers of participants who met each of these exclusion criteria were 4 and 9, respectively, resulting in a final sample of $N = 94$ (26 women, 67 men, and 1 person who chose another option; $M_{age} = 27.9$, $SD = 4.5$; 60.6% White, 13.8% Asian or Pacific Islander, 8.5% Black or African American, 7.4% Hispanic or Latino only, 2.1% American Indian or Alaskan, 4.3% mixed race or multiracial).

A sensitivity power analysis in G*Power ($\alpha = .05$; Faul et al., 2007) indicated that this sample size provided 80% power to detect a correlation of $r = .28$ (H1) and a difference between conditions of $d = 0.58$ (H2), and 60% power to detect a correlation of $r = .22$ and a difference of

d = 0.46. For reference, the median effect size in social psychology has been estimated at r = .21, or approximately d = 0.43 (Richard et al., 2003).

Procedure. Participants first completed a brief prescreen in which they indicated their age, gender, and whether they were primarily attracted to men or women. Only participants who were between 18 and 35 years old and primarily attracted to women were able to proceed. Next, participants saw the following instructions, adapted from Eastwick et al. (2019):

> In this study you will evaluate a series of faces that vary (from 0 to 100) in a characteristic called Reditry. Please pay careful attention to the information you see in this study, because we will ask you questions about it later on. Try to get an idea of your likes and dislikes, as well as how much Reditry each person has.

Participants then saw a series of 24 female faces, each presented along with its level of Reditry, and rated their romantic liking for each pictured person. After the trials, participants completed a measure of their overall summarized preference for Reditry. Lastly, after seeing another survey unrelated to the current research questions, participants completed the attention check and a demographic survey.

Materials and measures.

Stimuli. We selected 48 White female faces from the CFD (Ma et al., 2015). To manipulate average likeability, we divided the faces into two sets of 24 faces that varied similarly in babyfacedness (according to the norming-data ratings in Ma et al., 2015) and that differed only in how likeable they were on average. Likeability of these faces was rated in a previously published sample (Eastwick & Smith, 2018), in which 677 participants who were primarily attracted to women evaluated each face on a measure of romantic likeability using 1-7 rating scales. The faces we chose for the high likeability condition had a mean of M = 3.11 (SD = 0.60) on this scale, and the faces we chose for the low likeability condition had a mean of M = 2.10 (SD = 0.59). To avoid unintentionally manipulating the strength of the association between

babyfacedness and likeability, we ensured that the correlations between the Ma et al. (2015) ratings of babyfacedness and the Eastwick and Smith (2018) ratings of likeability were similar across conditions (r = .24 in both conditions); we also checked that this correlation was similar to the correlation between babyfacedness and likeability in the full population of White female faces in the CFD (r = .28). Finally, we inspected the scatterplot between these two variables in both conditions to ensure that they only differed in mean likeability; for example, the means, SDs, and ranges of babyfacedness were as similar as possible across conditions, the SDs and ranges of likeability were as similar as possible across conditions, and neither condition exhibited odd distributional properties (see Figure 1.2).

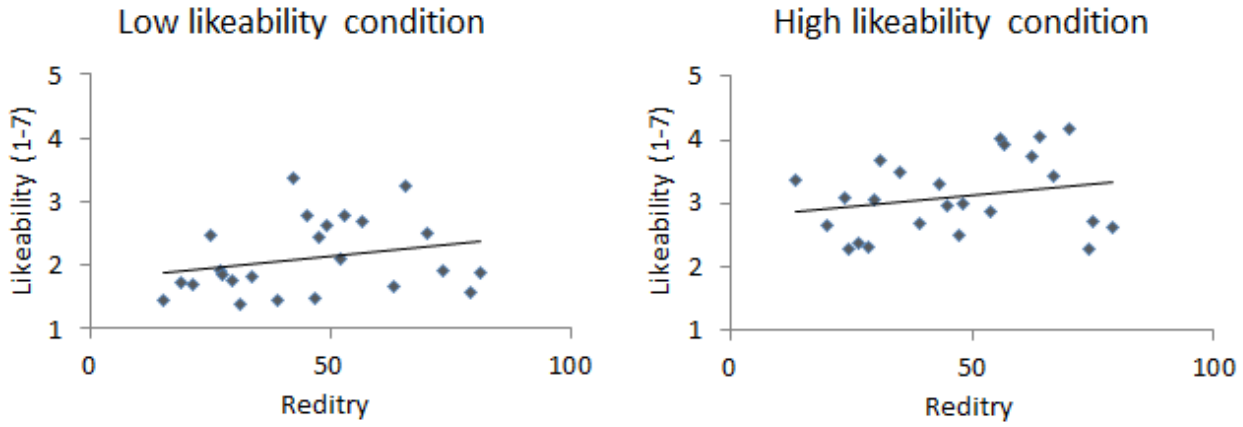

Figure 1.2. Scatterplots of the stimuli used for Study 1; each dot represents a face target. Notice that the correlation between pretest ratings of likeability and Reditry is the same in both conditions (i.e., the slopes of the trend lines were the same), whereas the average likeability is higher in the high (vs. low) likeability condition (i.e., the intercept of the trend line in the high vs. low likeability condition was higher).

After creating the two sets of faces, we rescaled the CFD rating of each face's babyfacedness to a value ranging from 0–100 and presented it to participants as the Reditry value of that face.

Functional preference measure. Following Wood and Brumbaugh's (2009) method, we measured participants' functional preferences for Reditry as the association between the level of

Reditry in the 24 targets with participants' experienced liking for those targets. On each screen, participants saw one target accompanied by the Reditry value of that face. They rated their experienced romantic liking for each target in response to the prompt "To what extent are you romantically interested in this person?" on a 9-point Likert-type scale (from -4 = *strongly dislike* to 4 = *strongly like*).[2] Presentation order of the targets was randomized.

Participants' functional preferences were calculated following Wood and Brumbaugh's procedures: First, each participant's romantic interest ratings were rescaled to a percentage-of-maximum-possible (POMP; Cohen et al., 1999) metric ranging from 0 to 100, such that 0 indicated the scale floor (*strongly dislike*) and 100 indicated the scale ceiling (*strongly like*).[3] Next, the POMP-rescaled ratings were regressed onto the levels of Reditry. Finally, the standardized regression coefficients from the regression models, akin to within-person slopes in linear mixed models, were *r*-to-*z* transformed (Fisher, 1925) to normalize the distributions for analysis. Each transformed regression coefficient represents a participant's own functional preference for a given attribute.

Summarized preference measure. After participants experienced their liking for all 24 faces, we measured their overall summarized preferences for Reditry with two items: "How much do you like Reditry in a romantic partner?" (from -4 = *strongly dislike* to 4 = *strongly like*) and "How desirable to you is Reditry in a romantic partner?" (from -4 = *extremely undesirable* to

[2] We used the term "romantic interest" rather than simply "liking" to ensure correspondence between the functional preference measure and the summarized preference measure, such that both assessed liking for Reditry in a romantic context (Ajzen & Fishbein, 1977; Ledgerwood et al., 2018); if we simply asked participants to report "liking," they might report liking for the targets as potential friends rather than as potential romantic partners. As measures of romantic evaluation, the terms "interest" and "liking" are interchangeable. In Study S2 (reported in the supplemental materials), items assessing romantic interest ("to what extent are you romantically interested…") and romantic liking ("to what extent do you romantically like…") were strongly associated, $\beta = .93$, 95% CI [.92, .94].

[3] Note that because we ran a standardized regression after the POMP transformation, the end result is the same as that without the POMP transformation. We followed this procedure to be consistent with previous research (e.g., Wood & Brumbaugh, 2009).

4 = *extremely desirable*). We averaged ratings on these two items to form an index of summarized preferences for Reditry (α = .92).

Winograd-like schema. We included an attention check that involved text interpretation to filter out bots and mindlessly responding participants, based on the structure of a Winograd schema (used to assess human-like reasoning; Levesque et al., 2011). Participants saw a short story: "Santa Claus is on vacation, and he goes to a beautiful beach on the Brazilian coast. He realizes he has forgotten sunscreen and wonders how he can protect his skin. Luckily, a young kid nearby understands the situation right away. As he wants to receive a nice gift for Christmas, he lends him a beach umbrella." Next, they answered two open-ended questions about the story ("Who receives the beach umbrella?" and "What does the kid hope will happen in December?"). Participants were excluded if they gave nonsensical answers (e.g., "brazilian"), as coded by a researcher blind to the study results.

Results

Manipulation check. We checked whether the manipulation of average target likeability successfully influenced the amount of liking that participants experienced when learning about their preferences. Our manipulation of average target likeability was successful: On average, participants in the high likeability condition experienced more liking for the faces they saw (M = -0.53, SD = 1.23) than participants in the low likeability condition (M = -1.67, SD = 1.35), $t(92) = 4.29$, $p < .001$, $d = 0.95$, 95% CI [0.52, 1.37].[4] Note that in general, romantic liking for these faces was on the lower side of the scale, which presumably reflects the fact that the CFD was designed to provide carefully controlled experimental stimuli (e.g., neutral expressions, minimal to no makeup) rather than to attract romantic partners.

[4] For this and all subsequent *t*-tests, we report Student's *t*-test for ease of interpretation; Welch's *t*-test yielded identical conclusions in all cases.

Functional preferences for Reditry. Although we took care to ensure that the correlation between Reditry and pretest ratings of face likeability were similar across conditions, it is still possible that our manipulation of average likeability could unintentionally influence participants' experienced functional preferences for Reditry. Thus, it was important to assess whether participants' experienced functional preferences for Reditry differed between the two conditions. Functional preferences were very similar across the two conditions (M = 0.24, SD = 0.26 vs. M = 0.20, SD = 0.15 for the high and low likeability conditions, respectively), $t(92)$ = 0.91, p = .365, d = 0.19, 95% CI [-0.22, 0.59], confirming that our manipulation of average target likeability did not affect participants' functional preferences for Reditry.

Main analyses. After confirming that our manipulation was successful at influencing average liking but not functional preferences for Reditry, we proceeded to our main analyses. First, we tested whether functional and summarized preferences were correlated (H1).[5] The correlation between functional and summarized preferences was r = .11, p = .309, 95% CI [-.10, .30]. Thus, the significance test provided no evidence that functional and summarized preferences were related in this learning context. The CI was however compatible with a relatively broad range of correlations, suggesting additional data would be informative.

Next, we tested whether average likeability of the targets influenced summarized preferences for Reditry (H2). Indeed, participants inferred stronger summarized preferences for Reditry in the high versus low likeability conditions (M = 0.18, SD = 1.74 vs. M = -0.89, SD = 1.84), $t(92)$ = 2.88, p = .005, d = 0.60, 95% CI [0.18, 1.01]. In other words, participants inferred that they liked Reditry substantially more when they learned about their preference in a context

[5] We did not plan to test H1 in the pre-analysis plans for Studies 1, 2, and 4 because it did not occur to us to include H1 at the time. We did record a plan to test H1 in the pre-analysis plan for Study 3, which was conducted after the other three studies.

with high (vs. low) likeability targets. Again, the CI was compatible with a broad range of effect sizes, suggesting additional data would be informative.

Discussion

The results of our first study suggest that when participants formed summarized preferences for an attribute for the first time, they based their summarized preferences on the average liking they experienced in the learning context but not their functional preferences for the attribute. These results provide initial evidence for the hypothesis that average likeability can influence summarized preference judgments (H2), but no evidence that functional and summarized preferences are themselves related (H1). In other words, people might form summarized preferences for an attribute that draw on seemingly incidental aspects of the learning context but not (or only weakly) on their actual functional preferences for that attribute.

This pattern of results begs the question: If summarized preferences are not necessarily related to functional preferences and if summarized preferences can be uniquely influenced by incidental features of the learning context, then is there any reason for researchers to study summarized preferences? As noted in the introduction, some researchers have argued that summarized preferences are simply crude proxies for functional preferences (Eastwick & Finkel, 2008; Brumbaugh & Wood, 2013), and our Study 1 results could be interpreted as consistent with this idea.

We see two reasons to pause before accepting this conclusion. First, it is possible that the lack of support for H1 in our first study simply reflects a Type II error. For example, our Study 1 sample size provided only 60% power to detect a correlation of $r = .22$, as discussed above. If the true correlation between functional and summarized preferences is relatively small, we may have been underpowered to detect it. Second, even if people can form summarized preferences

without drawing from their experienced functional preferences at all, summarized preferences may have some predictive power. In particular, one purpose of summarized preferences may be that they enable people to select into situations at a distance, based on socially acquired knowledge. Because summarized preferences are abstract ideas about likes and dislikes that are not tethered to any particular circumstance, they should be particularly useful when people are deciding on situations they have yet to encounter, without having to first experience those situations directly. In Study 2, we began to probe the possibility that summarized preferences can predict some interesting downstream consequences by including an additional item measuring situation selection at a distance.

Study 2

Study 2 sought to replicate and extend Study 1 in several ways. First, we created a new version of our paradigm with male faces rather than female faces, both to verify that our Study 1 results generalized across faces of both sexes and to disentangle two possible explanations for our Study 1 results. One possible explanation for the effect of average likeability on summarized preferences is the outcome density bias, as described in the introduction. However, it may also be possible to explain these results using a feature positive-effect account (Fazio et al., 1982; Jenkins & Sainsbury, 1970; Newman et al., 1980; Ward & Jenkins, 1965). Recall that in Study 1, participants displayed a positive functional preference for Reditry: On average, participants experienced greater liking for the female faces as babyfacedness increased. This positive functional preference meant that participants in the high (vs. low) likeability condition experienced more instances where they liked a high Reditry face. Insofar as people focus more on what happens in the presence rather than the absence of the feature (i.e., high Reditry), it seems possible that participants in the high (vs. low) likeability condition inferred a stronger

preference for Reditry simply because they noticed more instances in which they liked high Reditry faces. However, in a context where functional preferences are neutral (i.e., near-zero), the feature-positive "high Reditry, high likeability" faces should be equally common in the two conditions. Thus, if the same pattern of results were to appear when functional preferences are neutral, it would suggest that outcome density bias is a more likely mechanism than the feature positive effect. Because functional preferences for babyfacedness in male faces are near zero (r = .01 for White male faces in the CFD), using male faces allowed us to disentangle these two possible accounts. We hypothesized that average likeability would influence summarized preferences for Reditry (H2), even when functional preferences for Reditry were neutral.

Second, we also began to probe the possibility that summarized preferences predict situation selection at a distance. By means of socially acquired knowledge, humans have a profound ability to learn, communicate, and make decisions about situations at a distance before actually entering and personally experiencing them. Ancestrally, a hunter could decide which fields to visit based on someone's description of the characteristics of available prey; in modern times, a person can decide whether to try a new bar based on reviews that describe the patrons as particularly fun-loving or attractive. In the realm of online dating, platforms like The League and Sapio tout the high intelligence of their memberships (Murdoch, 2017), and people can decide whether to sign up for these websites based on socially acquired knowledge (e.g., verbal descriptions provided by others rather than their own direct experiences).

Theory and research suggest that abstract representations provide a crucial cognitive toolkit that humans can use to make future plans and navigate decision-making at a distance (Gilead et al., 2020; Fujita, 2011; Hofmann & Kotabe, 2012; Leary & Buttermore, 2003; Soderberg et al., 2015; Trope & Liberman, 2010; Wakslak et al., 2008). Importantly,

summarized preferences are relatively abstract: They reflect people's generalized evaluations of a trait, abstracted away from any one particular target or experience (Eastwick et al., 2019; Ledgerwood et al., 2018; Ledgerwood et al., 2020). Thus, it follows that summarized preferences, like other abstract representations, may enable people to make decisions about situations they have not yet directly experienced (H3). To begin probing this possibility, we added a new item to measure participants' interest in a dating website that was described as providing access to partners high in Reditry (i.e., a relevant situation that participants learned about through socially acquired knowledge—a verbal description—rather than direct experience). We hypothesized that summarized preferences for Reditry would predict participants' interest in joining this described website (H3).

Method

Participants. One hundred and eighty-four participants completed the study online through Mturk. Participants were randomly assigned to one of two conditions (low average likeability vs. high average likeability). An *a priori* power calculation in G*Power (Faul et al., 2007) suggested that to have 95% power to detect the effect size of $d = .60$ observed for our focal test of H2 in Study 1, we would need a total sample size of 148. We decided *a priori* to target a sample size of 170 with the goal of having at least $N = 150$ after planned exclusions; our actual total sample reached $N = 180$ because our survey software failed to count 10 participants with usable data who did not click to the last page of the survey. We used the same *a priori* exclusion criteria from Study 1. In this study, six participants gave the same ratings to all targets, and seven participants failed the attention check, resulting in a final sample of $N = 167$ (139 women, 21 men, and 7 people who chose another option; $M_{age} = 27.4$, $SD = 4.8$; 71.3% White, 7.2% Asian

or Pacific Islander, 7.2% Black or African American, 10.2% Hispanic or Latino only, 4.0% mixed race or multiracial, 4% reported a different identity or preferred not to answer).

Procedure. The procedure was identical to Study 1 except for two changes: (1) We used male faces instead of female faces and (2) we added an additional dependent measure as a first attempt to assess situation selection at a distance.

New materials and measures.

Stimuli. Similar to Study 1, we selected 48 White male faces from the CFD (Ma et al., 2015) and divided them into two sets of 24 faces that varied similarly in babyfacedness and differed only in how likeable they were on average. Based on ratings provided by 665 pretest participants who were primarily attracted to men (Eastwick & Smith, 2018), the average likeability of faces in the high likeability condition was M = 2.76 (SD = 0.59), and the average likeability of faces in the low likeability condition was M = 1.83 (SD = 0.38). We again ensured that the correlation between pretest ratings of babyfacedness and likeability was similar across conditions (r = .05 in both conditions) and reflected the actual correlation in the full population of White male faces in the CFD (r =.01). Again, we inspected the scatterplot between these two variables in both conditions and compared the descriptives to ensure that they only differed in mean likeability (see Figure 1.3).

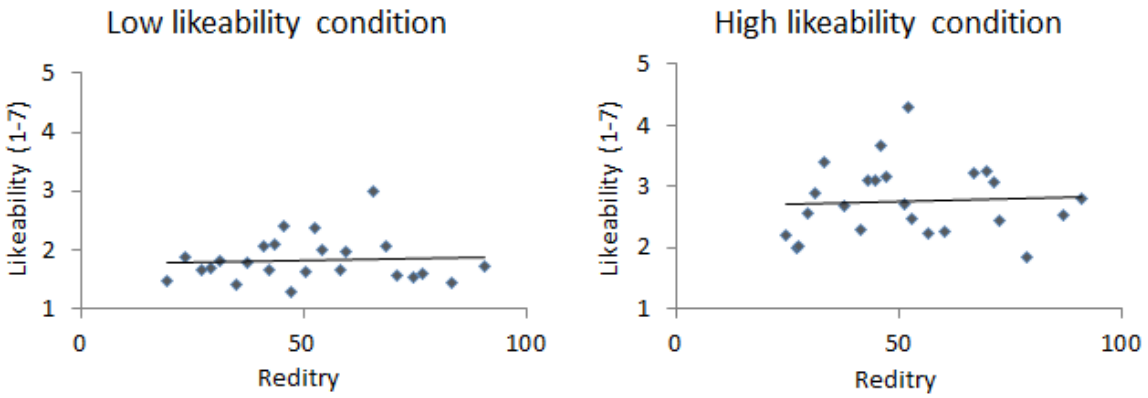

Figure 1.3. Scatterplots of the stimuli used for Study 2; each dot represents a face target. Notice that the correlation between pretest ratings of likeability and Reditry was neutral in both conditions (i.e., the slopes of the trend lines were the same), whereas the average likeability was higher in the high (vs. low) likeability condition (i.e., the intercept of the trend line in the high vs. low likeability condition was higher).

***Measure of situation selection at a distance*.** After completing the same functional and summarized preference measures used in Study 1, participants read a description of a situation that would provide them with access to potential partners high in Reditry: "Imagine that you are single and looking for a romantic partner. Imagine also that there is a dating website designed for people looking for partners high in Reditry. If you joined this website, you would have access to potential partners who are in the top 30% of Reditry." We asked participants how interested they would be in this website that would only include partners high in Reditry. Participants rated their interest on a 9-point Likert-type scale (1 = *not at all interested* to 9 = *very interested*).

Results

Manipulation check. Our manipulation of average target likeability was successful: On average, participants in the high likeability condition liked the faces they saw more (M = -0.80, SD = 1.23) than participants the low likeability condition (M = -1.87, SD = 1.37), $t(165) = 5.32$, $p < .001$, $d = 0.82$, 95% CI [0.50, 1.14].

Functional preferences for Reditry. We compared functional preferences for Reditry across the two conditions to check whether our manipulation of average target likeability unintentionally influenced average functional preferences for Reditry. Functional preferences were very similar across the two conditions (M = 0.12, SD = 0.26 vs. M = 0.10, SD = 0.24), $t(165) = 0.53$, $p = .599$, $d = 0.08$, 95% CI [-0.22, 0.38].

Main analyses. After confirming that our manipulation was successful at influencing average liking but not average functional preferences for Reditry, we proceeded to our main

analyses. First, we tested whether functional and summarized preferences were correlated (H1). The correlation between functional and summarized preferences was r = .09, p = .259, 95% CI = [-.06, .24]. As in Study 1, the significance test provided no evidence that functional and summarized preferences were related. The effect size estimate was similar to that obtained in Study 1, with a slightly narrower confidence interval (reflecting the greater precision afforded by the larger sample size of this study).

Next, we tested whether average likeability of the targets influenced summarized preferences for Reditry (H2). Replicating Study 1, participants inferred stronger summarized preferences for Reditry in the high versus low likeability conditions (M = -0.34, SD = 1.92 vs. M = -1.00, SD = 1.91), $t(165) = 2.23$, p = .027, d = 0.34, 95% CI [0.04, 0.65].

Finally, we tested whether summarized preferences for Reditry predicted situation selection at a distance (H3) by regressing interest in joining the dating website on participants' summarized preferences. Summarized preferences significantly predicted interest in the website, b = 0.62, SE = 0.08, p < .001, r = .53, 95% CI [.42, .63], providing initial evidence that summarized preferences might predict situation selection at a distance. Interestingly, functional preferences did not predict interest in joining the website, b = 0.93, SE = 0.69, p = .180, r = .10, 95% CI [-.04, .25]; we test this effect with stronger methods in Studies 3 and 4.

Discussion

The results of our second study replicated and extended Study 1, providing more evidence that when participants formed summarized preferences for an attribute for the first time, they based their summarized preferences on the average liking they experienced in the learning context (H2). Importantly, our manipulation of average likeability influenced participants' summarized preferences not only when average functional preferences for Reditry

were positive (for female faces, in Study 1), but also when average functional preferences for Reditry were neutral (for male faces, in Study 2). This pattern of results is consistent with outcome density rather than feature positivity as the underlying mechanism for the results observed in Study 2, although we cannot conclusively rule feature positivity out as an explanation for Study 1, given that Study 1 used a different population of participants (i.e., participants primarily attracted to women rather than men). In summary, then, the results of these studies provide support for the striking conclusion that people's summarized preferences for traits can be informed by seemingly incidental aspects of the context in which they learn about those preferences.

In both Study 1 and Study 2, the effect size estimates for the correlation between functional and summarized preferences were similar and quite small (r = .11 and r = .09, respectively), and significance testing in each individual sample provided no evidence for the hypothesis that summarized preferences would relate to functional preferences (H1). On the other hand, these estimates are consistent with the range of r = .02–.38 observed for a variety of traits in previous large-scale studies of preferences for traits in faces (e.g., Brumbaugh & Wood, 2013). Taken together, the results so far suggest that participants' summarized preferences for Reditry were probably weakly informed by their functional preferences for Reditry.

Perhaps most intriguingly, Study 2 provides a first hint that summarized preferences—even when only weakly based on functional preferences and when influenced by incidental contextual inputs—may still predict important outcomes. Specifically, participants' summarized preferences for Reditry predicted their interest in joining a dating website for high-Reditry partners (H3). Thus, it seems possible that even when functional and summarized preferences are only weakly related, summarized preferences might have important predictive power.

Study 3

As noted earlier, scholars studying attribute preferences in the context of human mating have tended to assume either that summarized and functional preferences can be measured interchangeably (e.g., Gerlach et al., 2019; see Ledgerwood et al., 2018, for a review), or that functional preferences are superior measures and should be assessed whenever possible (e.g., Eastwick & Finkel, 2008, Wood & Brumbaugh, 2009). In contrast to both views, the current data suggest that summarized preferences may have some unique consequences. That is, summarized preferences may be useful for situation selection at a distance, when people rely on socially acquired knowledge rather than direct experience to guide their decisions about which situations to enter (H3).

Of course, one might wonder whether our Study 2 results truly show a unique consequence of summarized preferences, or whether functional preferences simply did not predict situation selection at a distance because our measure of functional preferences was a poor measure that in fact would not predict anything. In contrast, consistent with broad principles of compatibility and matching (Azjen & Fishbein, 1977; Fujita et. al., 2008; Lee et. al., 2010) as well as how abstract mental tools are specifically recruited to support action at a distance (Trope et al., 2021), we predict that whereas summarized preferences should predict situation selection at a distance, functional preferences should predict situation selection with experience (i.e., a decision to enter a situation that participants have had an opportunity to sample). That is, once people have sampled targets from a novel situation (e.g., previewing other users on a dating website), they will (re-)experience their functional preferences during the sampling process and use those preferences to decide whether to enter the situation. For example, people can sometimes see photographs of other users on a dating website or sign up for a free trial before

deciding which dating platform to use. Once again, market researchers are interested in predicting how trial periods affect consumer purchasing decisions (e.g., Lee & Tan, 2013; van der Heijden et al., 2003).

Although not the central focus of our hypotheses, we should expect that functional preferences weakly (or do not) predict situation selection at a distance and that summarized preferences weakly (or do not) predict situation selection with experience. These predictions similarly draw from the principles of compatibility and matching: Summarized and functional preferences should be less relevant and predictive when they do not match the decisions that they support. When deciding whether to select into situations at a distance, people do not have access to their functional preferences as evaluative guides (which require that people directly experience those situations), and thus functional preferences could not guide those decisions. In contrast, when deciding on situations that people can sample, the relevance of summarized preferences as an evaluative guide diminishes in the face of functional preferences: People no longer need their abstract ideas about liking when they can directly recruit their experiences of liking for decision-making. In other words, to the extent that summarized preferences represent an abstract evaluative tool that people can use to make decisions at a distance, we expect that people will use them specifically for decision-making at a distance (see e.g., Ledgerwood et al., 2010; Trope et al., 2021, for similar reasoning). Therefore, to the extent that summarized and functional preferences are weakly correlated, the predictive power of summarized and functional preferences should be dissociable.

In Study 3, we set out to test the predictive power of existing summarized and functional preferences in the context of online dating, where people often have to weigh their interest in different dating websites that may offer access to different pools of partners. We tested both our

key hypothesis that summarized preferences would primarily predict situation selection at a distance (H3), as well as the corresponding hypothesis for functional preferences:

> H4: Functional preferences will predict situation selection when people can directly sample a situation.

As in Studies 1 and 2, we also tested the hypothesis that summarized preferences would correlate with functional preferences (H1).

Following the measurement of summarized and functional preferences, we introduced participants to dating websites that would provide access to members who are high in Reditry. We designed our websites so that some provided participants with an opportunity to sample targets from the website (by viewing photographs of users), whereas another did not provide participants with such an opportunity (participants simply read descriptions of the website). We tested how summarized and functional preferences would respectively predict participants' website selection at a distance and website selection with experience. We preregistered our pre-analysis plan on OSF at: https://osf.io/c8p5a/?view_only=f162cf7b9b2941809c2343d230ba97a6.

Method

Participants and power. Five hundred and eighty-six participants completed the study online through MTurk. As in Studies 1–2, we limited the range of participants to 18–35 years old and primarily attracted to males. In Study 2, the correlation between functional and summarized preferences was $r = .09$, $p = .259$, 95% CI = [-.06, .24]. We planned to power this study to obtain a stable estimate of this correlation. Based on Schönbrodt and Perugini (2013), we need at least 470 participants to reach a corridor of stability of width = .10 in a 95% confidence interval. We decided to collect a larger sample size to have at least $N = 550$ after exclusions to provide a stable estimate of the effect size. We used the same *a priori* exclusion criteria from Studies 1–2.

In this study, 12 participants gave an identical response to all photographs and 4 failed the attention check, resulting in a final $N = 570$ (519 women, 42 men, and 9 a different identity; $M_{age} = 27.9$, $SD = 4.5$; 63.2% White, 10.5% Asian or Pacific Islander, 8.2% Black or African American, 7.4% Hispanic or Latino only, .05% American Indian or Alaskan, 8.8% mixed race or multiracial, 1.4% reported a different identity or preferred not to answer).

Procedure. All participants completed measures of functional and summarized preferences for Reditry, followed by the attention check. Next, we told participants that the research team was developing a series of dating websites and we asked them to indicate their interest in those dating websites. Our situation selection at a distance measure was identical to the dependent measure of Study 2. In this measure, we described a dating website as designed "for people looking for partners high in Reditry." Participants learned that if they joined this website, they would "have access to potential partners who are in the top 30% of Reditry," and they rated their interest in joining. In addition, we assessed participants' interest in selecting into situations that they had an opportunity to directly experience. In this situation selection with experience measure, participants learned about two websites, Website A and Website B, and they had the opportunity to sample these websites via an ostensible screenshot of each website presented side by side, one with faces higher in Reditry on average and the other with faces lower in Reditry on average (Figure 1.4). We counterbalanced the order of the two situation-selection dependent measures (i.e., at a distance vs. with experience) across participants. Last, participants provided their demographic information.

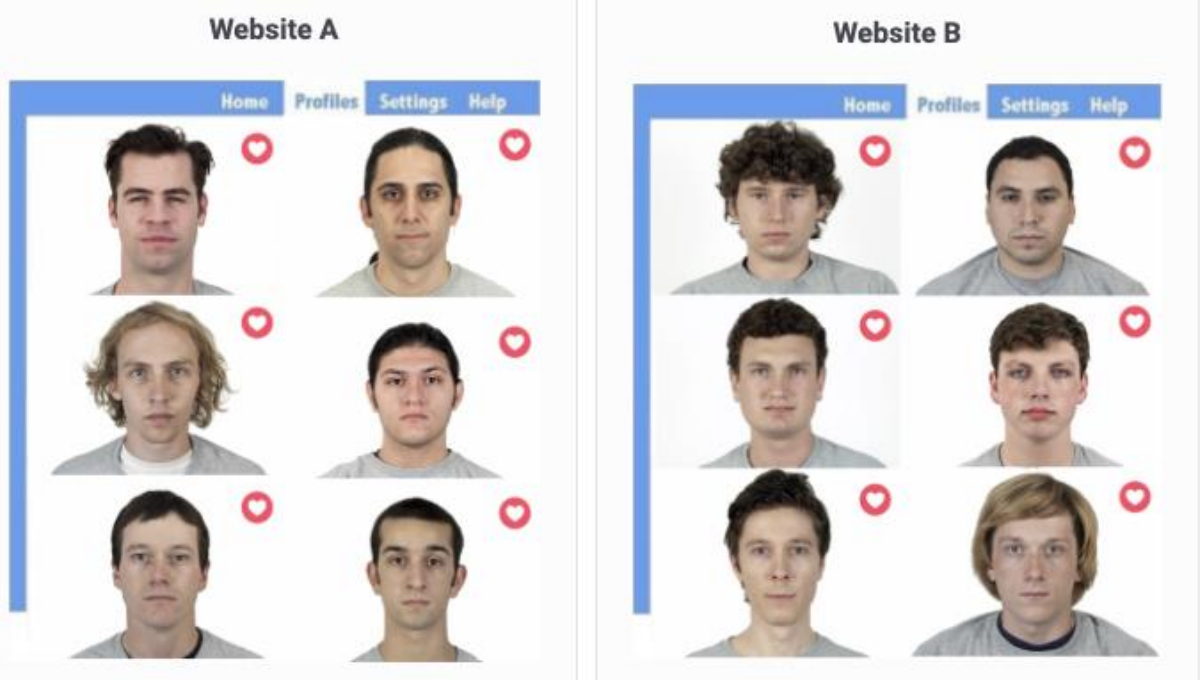

Figure 1.4. Stimuli used in Study 3 for the dependent measure of situation selection with experience. The screenshot of Website A presented photographs of six targets that had been rated lower in Reditry (babyfacedness), and the screenshot of Website B presented photographs of six targets that had been rated as higher in Reditry. The two websites appeared side by side on the same screen and participants selected their choice by clicking on one of the two screenshots.

New materials and measures.

Functional preference measure. Participants saw 40 White male faces from the CFD (Ma et al., 2015), one at a time. On each screen, participants saw one target accompanied by the Reditry value of that face. They rated their experienced romantic liking for each target in response to the prompt "To what extent are you romantically interested in this person?" on a 9-point Likert-type scale (from -4 = *strongly dislike* to 4 = *strongly like*). Participants' functional preferences for Reditry were calculated using the same procedure as in Studies 1 and 2.

***Measure of situation selection with experience*.** To assess participants' interest in entering a situation with (a) partners low in Reditry or (b) partners high in Reditry after having a chance to sample targets from those situations, we presented participants with screenshots of two dating websites and asked them to indicate which website they would choose to join. The first

website screenshot contained six photographs of targets that were lower in Reditry on average (M = 1.92, SD = 0.22); the second website screenshot contained six photographs of targets that were higher in Reditry on average (M = 3.36, SD = 0.42; see Figure 1.4). The websites were matched in attractiveness (M = 2.74, SD = 0.33 and M = 2.74, SD = 0.38, respectively).

Results

Hypothesis 1. The correlation between functional and summarized preferences for Reditry was r = .11, p = .008, 95% CI [.03, .19]. Notably, this point estimate is consistent with those observed in our previous studies (Study 1: r = .11, 95% CI [-.10, .30]; Study 2: r = .09, 95% CI [-.06, .24]), as well as large-sample studies in the literature (e.g., Brumbaugh & Wood, 2013; Eastwick & Smith, 2018; Wood & Brumbaugh, 2009), again suggesting that across studies, functional and summarized preferences were weakly correlated.

Hypotheses 3 and 4. Next, we tested our hypotheses that summarized preferences would predict situation selection at a distance (H3), whereas functional preferences would predict situation selection with experience (H4). For this relatively complex set of analyses, we followed our pre-analysis plan to constrain researcher degrees of freedom; all analyses reported below were preregistered unless explicitly noted in the text. Because multiple analytic approaches were possible with our data, we decided *a priori* to focus on the effect sizes and *p*-values from one focal approach (structural equation modeling [SEM], as described below), which would allow us to think about those *p*-values as diagnostic of the likelihood of a given statistical result (de Groot, 2014; Nosek et al., 2018), while also considering the consistency of the patterns across alternative analytic approaches (e.g., multiple regression). In other words, we planned to calibrate our confidence in our results based on both the extent to which focal *p*-values reached significance and on the extent to which similar patterns of effect sizes emerged across different

analytic approaches. We decided to focus primarily on the effect sizes and *p*-values from the SEM approach because estimates from latent variable models tend to be more accurate (less biased) than those from observed variable approaches and because an SEM approach helps avoid Type I error inflation in this multivariate context (Ledgerwood & Shrout, 2011; Wang & Eastwick, 2020). At the same time, any one estimate from SEM analyses using latent variables can be quite far from the true population parameter (Ledgerwood & Shrout, 2011), so looking for consistent patterns across multiple analytic approaches can be informative.

Our planned focal approach was therefore to use SEM to test the effect of a summarized preference on a dependent variable, while controlling for the functional preference, and vice-versa (e.g., testing the effect of summarized preference for Reditry on a dependent variable, controlling for functional preference for Reditry). In the SEM analysis, both dependent variables were simultaneously regressed on both predictors (i.e., summarized and functional preferences); the predictors were modelled as latent factors (Figure 1.5). Latent factors of summarized preferences had two indicators, whereas latent factors of functional preferences were indicated by fixing participants' functional preferences to the reliability of .70, as it provides a reasonable tradeoff between Type I error rate and power (Savalei, 2019). The SEM model provides a good fit of the data, $\chi^2(3) = 0.58$, $p = .901$, Comparative Fit Index (CFI) = 1.00, Tucker-Lewis Index (TLI) = 1.03, Root Mean Square Error of Approximation (RMSEA) = 0.00.

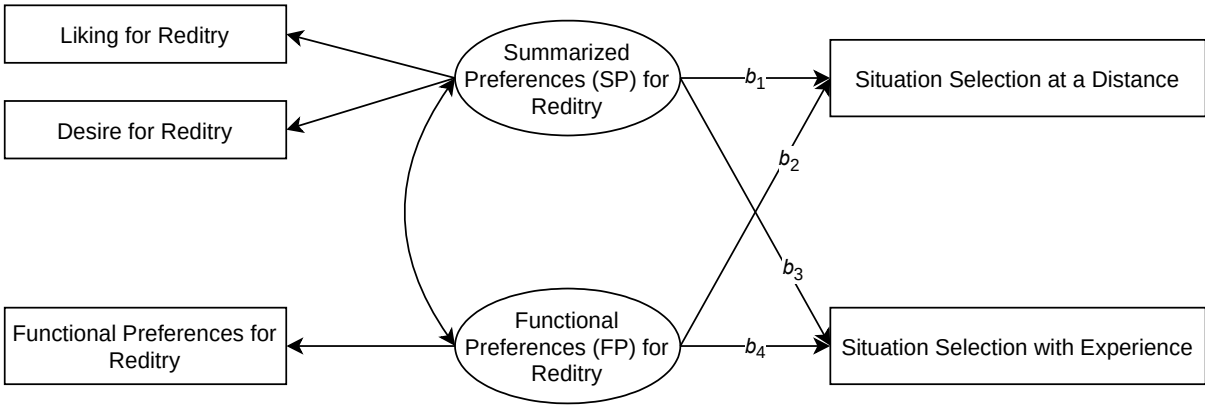

Figure 1.5. Diagram of the model of double dissociation in which dependent variables were regressed on attribute preferences as latent predictors. Key parameters showing the double dissociation pattern are denoted b_1–b_4 and reported in Table 1.1. For visual simplicity, residual (co)variances are not shown. Higher values on both situation selection variables indicate a tendency to select into higher Reditry situations.

We also planned to model the unique effects of summarized and functional preferences by conducting multiple regressions in which each dependent variable was regressed on both types of preferences and examining the effect size estimates provided by the partial regression coefficients of the predictors. Finally, we planned to conduct simple regressions as well (i.e., one preference predicting one dependent variable).

Table 1.1 presents the results from all three analytic approaches. Hypothesis 3 received support, just as in Study 2: Summarized preferences significantly predicted situation selection at a distance using all three approaches (with moderate-to-large effect sizes). Also as in Study 2, functional preferences did not predict situation selection at a distance especially strongly: Only one of the three approaches was significant, and effect sizes were considerably smaller than for summarized preferences. In our focal SEM approach, summarized preferences predicted situation selection at a distance more strongly than functional preferences ($b_1 > b_2$, $\Delta\chi^2(1) = 27.46$, $p < .001$; exploratory/not planned).[6] The effect size difference between summarized and functional preferences was more or less the same across all three approaches.

Table 1.1.

Summarized and Functional Preferences Predicting Primary Dependent Variables in Study 3.

Analytic Approaches	Predictor Type	Dependent Variables	*b* (SE)	*p*	β	*OR*	*r* [95% CI]
Structural	SP	SS_d	0.49 (0.05)	< .001	.41	-	.40 [.33, .47]
Equation	FP	SS_d	1.00 (0.55)	.066	.07	-	.07 [-.00, .14]
Models	SP	SS_e	0.02 (0.03)	.554	-	1.06	.02 [-.04, .07]

[6] To compare the relative strength of the effects, we conducted likelihood-ratio tests by comparing the exact fit of the original model, where the regression coefficients of summarized and functional preferences were freely estimated, with that of a model with equality constraints on those regression coefficients.

	FP	SS_e	1.26 (0.41)	.002	-	1.39	.09 [.03, .15]
Bivariate	SP	SS_d	0.48 (0.04)	< .001	.41	-	.41 [.33, .48]
Regression	FP	SS_d	1.27 (0.50)	.011	.11	-	.11 [.02, .19]
	SP	SS_e	0.05 (0.04)	.248	-	1.10	.03 [-.02, .07]
	FP	SS_e	1.57 (0.47)	< .001	-	1.34	.08 [.03, .13]
Multiple	SP	SS_d	0.46 (0.05)	< .001	.40	-	.40 [.32, .47]
Regression	FP	SS_d	0.75 (0.46)	.104	.06	-	.06 [-.01, .14]
	SP	SS_e	0.03 (0.05)	.453	-	1.07	.02 [-.03, .06]
	FP	SS_e	1.53 (0.48)	.001	-	1.33	.08 [.03, .13]

Note. SP = summarized preferences, FP = functional preferences, SS_d = situation selection at a distance, SS_e = situation selection with experience, OR = odds ratio. Unstandardized regression coefficients (*b*) for situation selection with experience are logit coefficients.

Hypothesis 4 also received support. Functional preferences predicted situation selection with experience using all three approaches with modest but nevertheless significant effect sizes. Summarized preferences did not significantly predict situation selection with experience using any of the three approaches, and effect sizes were approximately zero. In our focal SEM approach, functional preferences predicted situation selection with experience more strongly than summarized preferences, although this difference was not significant ($b_4 > b_3$, $\Delta\chi^2(1) = 3.20$, $p = .074$; exploratory/not planned). The effect size difference between summarized and functional preferences was more or less the same across all three approaches. In total, the pattern of results suggests that we can have a relatively high degree of confidence that both H3 and H4 (i.e., the double dissociation pattern) received support.

Discussion

The results of Study 3 suggested that although summarized and functional preferences may be only weakly related, both have predictive power. First, we observed correlations between summarized and functional preferences at levels very similar to those of preferences for Reditry in Studies 1–2 (i.e., $r = \sim.10$). Second, we observed a double dissociation between summarized and functional preferences, such that summarized preferences predicted situation selection at a distance (as when people read a description of a website; H3), whereas functional preferences

predicted situation selection with experience (H4). The nonpredicted paths (i.e., summarized preferences predicting situation selection with experience; functional preferences predicting situation selection at a distance) tended to be very small and not significant. In other words, summarized preferences seem to have predictive power when participants are considering a situation-selection decision in the abstract, whereas functional preferences seem to have predictive power when participants are selecting into a situation they have had a chance to sample. These results were similar across three analytic approaches (including when measurement error was taken into account with SEM), thus increasing our confidence in their robustness.

Our Reditry paradigm circumvented participants' pre-existing expectations and summarized preferences by requiring them to learn about a novel trait, thereby creating an ideal context in which to study how summarized preferences form in the first place (Studies 1–3). However, the high experimental control of this paradigm potentially comes at the cost of external validity. To better understand what summarized attribute preferences predict, it is also important to study existing preferences for familiar traits. In addition, to better understand how attribute preferences operate in the realm of romantic attraction, it is also important to move from carefully controlled stimuli like White CFD faces to more externally valid and diverse stimuli like real-world dating profiles. In the next study, we turn to a more externally valid paradigm to better illuminate the consequences of attribute preferences.

Study 4

In Study 4, we set out to test the predictive power of summarized and functional preferences for known, familiar attributes in the real-world context of online dating, where people often have to weigh their interest in different dating websites that may offer access to

different pools of partners. We again tested the hypotheses that summarized and functional preferences would be associated (H1), that summarized preferences would primarily predict situation selection at a distance (H3), and that functional preferences would predict situation selection when people can directly sample a situation (H4). As in Study 3, we designed these situation-selection DVs to mimic real-life online dating contexts where people can select websites either at a distance (as when people simply read a description of a website or learn about it from friends) or after sampling the situation (as when people see photographs of other users on the website or sign up for a free trial).

We selected two focal attributes, intelligence and confidence, and measured participants' summarized and functional preferences for the two attributes in potential romantic partners. We used intelligence and confidence as our focal attributes because they can be readily inferred from faces (Oosterhof & Todorov, 2008), allowing us to assess functional preferences following a photograph-evaluation procedure employed in past research on partner preferences (Brumbaugh & Wood, 2013; Eastwick & Smith, 2018; Wood & Brumbaugh, 2009). Using two focal attributes rather than only one also provided us an opportunity to check whether the pattern of results would replicate across different attributes.

Method

The preregistration is publicly available on the Open Science Framework at: https://osf.io/tqfvc/?view_only=48d41b4de11245a78fc64aeb330c15cd.

Participants. Six hundred and eighty-four participants completed the study online through MTurk (see Power Analyses section below for a discussion of our sample size determination). As in Studies 1–2, we limited the age range of participants to 18–35 years old; this time, we included both participants who were primarily attracted to men and those who were

primarily attracted to women in a single study. We preregistered four *a priori* exclusion criteria: We would exclude participants who (1) gave an identical rating to all faces presented for measurement of functional preferences, (2) gave an identical rating to all attributes presented for measurement of summarized preferences, (3) provided a response other than male or female to the question asking about their gender (to maintain comparability to other similar studies; e.g., Eastwick & Smith, 2018), and/or (4) failed the attention check presented before the measurement of our dependent variables. In this sample, the number of participants who met each of these exclusion criteria were 9, 3, 5, and 115, respectively. Excluding these participants resulted in a final $N = 555$ (337 women, 218 men; $M_{age} = 28.9$, $SD = 4.0$; 71.5% White, 12.6% Black/African American, 6.3% Asian or Pacific Islander, 1.4% Native American, 5.9% mixed race or multiracial, 2.2% reported a different race); note that some participants met more than one exclusion criterion.

Procedure. We asked participants to imagine that they were single and looking for a romantic partner. First, they indicated the sex to which they were primarily romantically attracted, which determined the sex of the potential partners presented to each participant throughout the rest of the study. Then, participants completed measures of (a) summarized preferences for intelligence and confidence and (b) functional preferences for intelligence and confidence (order of attributes and order of summarized versus functional measures were each randomized across participants). All participants then completed an attention check.

Next, participants saw the situation selection measures, which were similar to the ones used in Study 3. To assess participants' interest in selecting into situations at a distance, without experiencing or sampling any targets from those situations, we presented a pair of dating websites and described one dating website as designed "for people looking for smart partners."

Participants learned that if they joined this website, they would "have access to potential partners who are in the top 30% of intelligence." We described the other dating website as designed "for people looking for self-assured partners." Participants learned that if they joined this website, they would "have access to potential partners who are in the top 30% of confidence." We then measured participants' interest in these two websites (the order in which participants indicated interest for these two websites was randomized).

To assess participants' interest in selecting into situations that they had an opportunity to directly sample, we presented participants with a different pair of dating websites, Website A and Website B. We gave participants the opportunity to sample these websites by showing them an ostensible screenshot of each website, which contained photographs of targets that were particularly high on one of the attributes. Because positive attributes (like intelligence and confidence) tend to be correlated in face impressions (e.g., Stolier et al., 2018), we selected photographs that were high on one attribute but not the other, so that participants' responses to a given website would indicate interest in a situation with higher levels of the attribute in question, rather than interest in a generically positive situation. Thus, the screenshot of Website A consisted of photographs of targets that were high on confidence but low on intelligence, and the screenshot of Website B consisted of photographs of targets that were high on intelligence but low on confidence. The screenshots were presented side by side. We then measured participants' interest in these two websites.

As in Study 3, we randomized the order of the situation-selection dependent variables. Last, participants provided their demographic information.

Materials and measures.

Summarized preference measure. To assess participants' existing summarized preferences for familiar traits, we presented them with a list of 16 traits and they rated the extent to which they desired each attribute in an ideal romantic partner on a 7-point Likert-type scale (from 1 = *not at all* to 7 = *a great deal*; adapted from Joel et al., 2017). Participants' summarized preference for intelligence was calculated as the mean of ratings for *intelligent*, *smart*, and *intellectually sharp* (α = .89), and participants' summarized preference for confidence was calculated as the mean of ratings for *confident* and *self-assured* (α = .79; see the supplemental materials for the full list of rated attributes and their descriptive statistics).[7]

Functional preference measure. To assess participants' functional preferences for intelligence and confidence, we adapted the same measures used in Studies 1–3 with one key change: To enhance the external validity of our study, we used photographs that we collected from actual dating profiles on a publicly accessible dating website (100 male targets, 100 female targets) rather than the carefully posed faces from the Chicago Face Database. We collected trait ratings for each target profile in an independent MTurk sample (N = 132; see Study S3 in the supplemental materials for details).[8] Our measure of liking for each target was also slightly different simply because the materials were designed by a different researcher: Participants rated the extent to which they experienced romantic desire (rather than "romantic interest") for each target, again on a 9-point Likert-type scale (from 1 = *not at all* to 9 = *a great deal*).[9] Functional

[7] We collected ratings on one additional item related to confidence ("charismatic"). Following our pre-analysis plan, we dropped the item from our calculation of summarized preference for confidence because including this item lowered the internal consistency of scale by more than $\Delta\alpha$ = .01. Including the item did not substantively change our results (e.g., no change of levels of significance, and no decline in the fit of our structural equation models; see the supplemental materials for details).

[8] Just as in real-life online dating contexts, we are agnostic of the "true" level of intelligence and confidence in our targets. Rather, levels of attributes are inferred from faces, and past research suggests that the focal attributes we measured elicit a high level of consensus: In other words, people agree on how intelligent and confident targets look (Oosterhof & Todorov, 2008).

[9] Items assessing "romantic interest," "romantic desire," and "romantic liking" can be viewed as interchangeable. In Study S2, romantic desire was strongly associated with both romantic interest, $\beta_{desire.interest}$ = .87, 95% CI [.86, .89], and romantic liking, $\beta_{liking.desire}$ = .84, 95% CI [.82, .86], all ps < .001 (see the supplemental materials for details).

preference for an attribute was calculated in the same way as Studies 1–3: Each participant's romantic desire ratings were rescaled to a POMP metric ranging from 0 to 100, such that 0 indicated the scale floor (*not at all*) and 100 indicated the scale ceiling (*a great deal*). Next, the POMP-rescaled ratings were regressed onto the levels of the attribute. Finally, the standardized regression coefficients from the regression models were *r*-to-*z* transformed. Each transformed regression coefficient represents a participant's own functional preference for a given attribute (α = .78 for intelligence and α = .83 for confidence).

Attention check. We again included an attention check to filter out inattentive participants, this time adapted from the standard instructional manipulation check (IMC; Oppenheimer et al., 2009). A paragraph embedded within the study procedure instructed participants to ignore a question that appeared underneath the paragraph and instead simply confirm that they had read the instructions.

Situation selection at a distance. To assess participants' interest in entering a not-yet-experienced situation with (a) highly intelligent partners or (b) highly confident partners, we adapted the situation selection item from Studies 2–3: Participants indicated how interested they were in the website described as providing access to partners in the top 30% of intelligence, and (in a separate question) how interested they were in the website providing access to partners in the top 30% of confidence on a 9-point Likert-type scale (from 1 = *not at all interested* to 9 = *very interested*).

Situation selection with experience. To assess participants' interest in entering a situation with (a) highly intelligent partners or (b) highly confident partners after having a chance to sample targets from those situations, we presented participants with screenshots of two websites and asked them to indicate which dating website they would choose to join. The first

website screenshot contained six photographs of targets that were relatively high on confidence (top 40% of our stimuli set) but middling on intelligence (bottom 40% of our stimuli set), whereas the second website screenshot contained six photographs of targets that were relatively high on intelligence (top 40% of our stimuli set) but middling on confidence (bottom 40% of our stimuli set; see Figure 1.6).

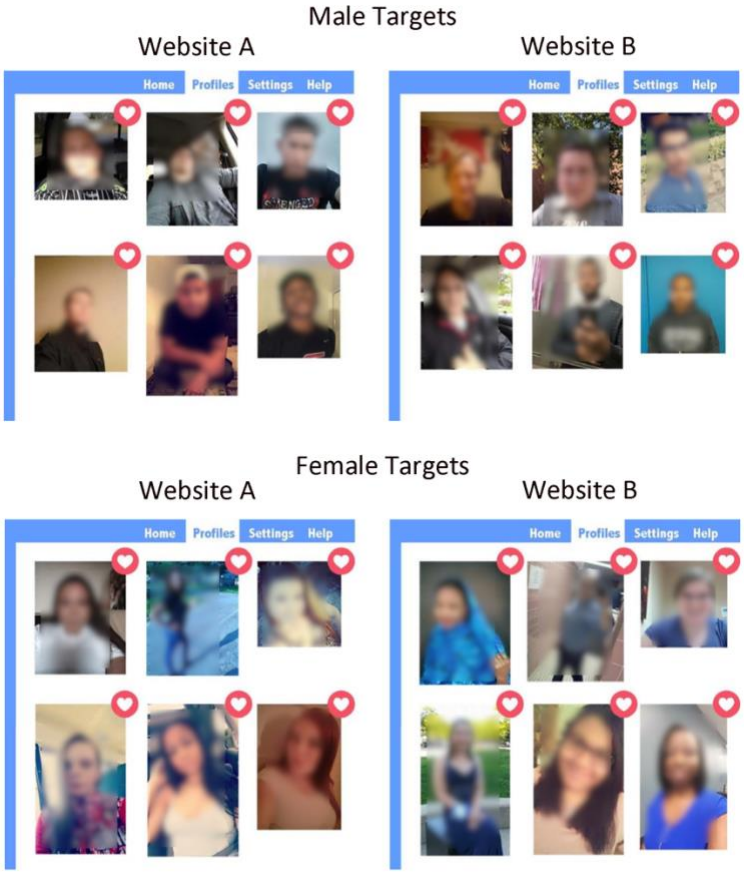

Figure 1.6. Stimuli used in Study 4 for the dependent measure of situation selection with experience. The screenshot of Website A presented photographs of six targets that had been rated in an independent sample (Study S3 in the supplemental materials) as relatively high on confidence but middling on intelligence, and the screenshot of Website B presented photographs of six targets that had been rated as relatively high on intelligence but middling on confidence. The two websites appeared side by side on the same screen and participants selected their choice by clicking on one of the two screenshots.

Secondary dependent measures. Recall our focal hypotheses: We expected that summarized preferences would predict situation selection at a distance (H3), whereas functional

preferences would predict situation selection with experience (H4). To explore the extent to which such results might be driven by the particular format of the primary dependent measures described above, we included two additional, secondary dependent measures after participants read about the websites that would provide access to (a) highly intelligent partners and (b) highly confident partners. First, to examine whether summarized preferences would only predict situation selection when the dependent measure focuses on one situation at a time (i.e., as in our primary *situation selection at a distance* measure), we included a measure that forced a tradeoff between the two situations: Participants indicated how interested they were in one website versus the other on a 9-point bipolar scale (1 = *the website that would only include intelligent partners*, 9 = *the website that would only include confident partners*). Second, to examine whether functional preferences would specifically predict a choice between two experienced situations (i.e., as in our primary *situation selection with experience* measure) or more broadly predict any kind of binary choice, we included a measure that asked participants to choose between the two described websites: Participants indicated which of the two described websites they would choose to join if both websites were available to them at the same price.

Power analyses for determining sample size. We determined our target sample size by running a series of power analyses using Monte Carlo simulations (Muthén & Muthén, 2002; Wang & Rhemtulla, 2021). We powered our study at 80% (with $\alpha = .05$) to detect the three quantities of interest that would be most difficult to detect in our design: (1) the effect of functional preferences for intelligence on choice between experienced websites in our planned structural equation model, controlling for functional preferences for confidence, (2) the effect of functional preferences for confidence on choice between experienced websites in the structural equation model, controlling for functional preferences for intelligence, and (3) level of model

misfit in the structural equation model. In power analyses for the first two effects, we used parameter estimates observed in a preliminary study (N = 332; β_1 = .13, β_2 = .18) to create the population model, from which we generated simulated data. In power analysis for the third effect, we followed the procedure described by MacCallum et al. (1996) by specifying the null hypothesis of close fit as H0: RMSEA = 0.05 and the alternative hypothesis of not-close fit as Ha: RMSEA = 0.10. These simulation-based power analyses showed that the minimum target sample size that would give us at least 80% power to detect all three effects was 535. We anticipated an exclusion rate of at least 15% based on a preliminary study and oversampled to ensure that we would have at least N = 535 for analysis. All power analyses were conducted in R using the 'lavaan' package (R Core Team, 2018; Rosseel, 2012).

Results

Hypothesis 1. We had no pre-analysis plan for testing whether functional preferences predict summarized preferences (H1); thus, we followed the same analysis used to test this hypothesis in the prior three studies. Note that the large sample size employed in this study exceeds Schönbrodt and Perugini's (2014) minimum recommendation for stable effect size estimates, and so we can have a relatively high degree of confidence in the stability of the observed correlations. The correlation between functional and summarized preferences for intelligence was r = .18, p < .001, 95% CI [.10, .26], and the correlation between functional and summarized preferences for confidence was r = .08, p = .045, 95% CI [.002, .17]. Notably, the CIs for both attributes were compatible with the CIs observed for Reditry in our previous studies (meta-analytic results for Studies 1-3 and Study S1 in the Supplemental Materials: N = 1046, r = .12, z = 3.74, p < .001, 95% CI [.06, .18]), again suggesting that across studies, functional and summarized preferences were weakly correlated.

Hypotheses 3 and 4: preregistered analyses. Next, we tested our hypotheses that summarized preferences would predict situation selection at a distance (H3), whereas functional preferences would predict situation selection with experience (H4). As in Study 3, we decided *a priori* to focus on the effect sizes and *p*-values from one focal approach (SEM), while also considering the consistency of the patterns across two alternative analytic approaches (bivariate and multiple regressions).

Our planned focal approach was to use SEM to test the effect of a summarized or functional preference for an attribute on a dependent variable, while controlling for the same type of preference for the other attribute (e.g., testing the effect of functional preference for confidence on a dependent variable, controlling for functional preference for intelligence; see Figure 1.7 for conceptual diagrams). In each SEM analysis, the dependent variable was simultaneously regressed on two predictors that were modelled as latent factors. Latent factors of summarized preferences were measured with each item as an indicator (i.e., *intelligent*, *smart*, and *intellectually sharp* for intelligence, and *confident* and *self-assured* for confidence). Latent factors of functional preferences were measured by randomly dividing the 100 target stimuli into four parcels, then calculating participants' functional preferences from each parcel as an indicator (following a random parceling approach; Little et al., 2002). Because the same parcels were used to calculate functional preferences for both intelligence and confidence, we allowed residual covariances of matching parcels (e.g., functional preferences for intelligence from parcel 1 and functional preferences for confidence from parcel 1) to be freely estimated.

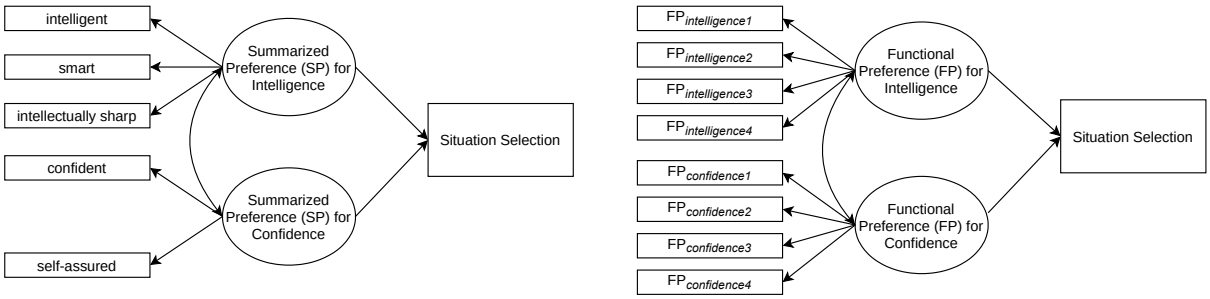

Figure 1.7. Conceptual diagrams of the planned structural equation models testing H3 and H4. Each situation selection dependent variable was simultaneously regressed onto summarized preferences (left panel) or functional preferences (right panel). For visual simplicity, residual (co)variances are not shown.

We further planned to interpret the results of our focal approach in the context of two other analytic approaches: bivariate and multiple regressions, which provide estimates that are conceptually akin to raw and semipartial correlations, respectively. For the bivariate regression approach, we planned to model the direct effects of summarized and functional preferences by regressing dependent variables on preference variables (as composite scores) and examining the effect size estimates as regression coefficients of the predictors. For the multiple regression approach, we planned to model the partial effects of summarized and functional preferences by conducting multiple regressions in which dependent variables were simultaneously regressed on the same type of preferences for the two attributes and examining the effect size estimates provided by the partial regression coefficients of the predictors.

***Primary situation-selection dependent measures*.** Our main analyses tested the extent to which summarized preferences and functional preferences each predicted situation selection at a distance (H3) and situation selection with experience (H4; Tables 1.2A and 1.2B). All models fit the data reasonably well, χ^2s = 1.23–265.30, CFIs = 0.90–1.00, TLIs = 0.84–1.00, RMSEAs = 0.02–0.14 (see supplement for details).

Table 1.2A.

Summarized and Functional Preferences Predicting Primary Dependent Variables for Intelligence in Study 4.

Analytic Approaches	Predictor Type	Dependent Variables	*b* (*SE*)	*p*	*β*	*OR*	*r* [95%CI]
Structural Equation Models	SP	SS at a distance	0.90 (0.13)	< .001	0.39	-	.32 [.24, .41]
	FP	SS at a distance	0.40 (0.18)	.026	0.17	-	.11 [.01, .21]
	SP	SS with experience	0.44 (0.13)	< .001	-	1.55	.12 [.05, .19]
	FP	SS with experience	1.06 (0.13)	< .001	-	2.87	.28 [.22, .34]
Bivariate Regression	SP	SS at a distance	0.73 (0.09)	< .001	0.33	-	.33 [.15, .50]
	FP	SS at a distance	1.06 (0.49)	.031	0.09	-	.09 [.008, .17]
	SP	SS with experience	0.13 (0.10)	.192	-	1.13	.03 [-.02, .09]
	FP	SS with experience	0.46 (0.48)	.339	-	1.59	.13 [-.13, .36]
Multiple Regression	SP	SS at a distance	0.75 (0.10)	< .001	0.34	-	.30 [.12, .47]
	FP	SS at a distance	1.36 (0.66)	.039	0.12	-	.09 [.004, .17]
	SP	SS with experience	0.32 (0.11)	.004	-	1.38	.09 [.03, .15]
	FP	SS with experience	5.48 (0.78)	< .001	-	239. 86	.83 [.74, .88]

Note. SP = summarized preferences, FP = functional preferences, SS = situation selection, OR = odds ratio. Unstandardized regression coefficients (*b*) for situation selection with experience are logit coefficients.

Table 1.2B.

Summarized and Functional Preferences Predicting Primary Dependent Variables for Confidence in Study 4.

Analytic Approaches	Predictor Type	Dependent Variables	*b* (*SE*)	*p*	*β*	*OR*	*r* [95%CI]
Structural Equation Models	SP	SS at a distance	0.86 (0.14)	< .001	0.38	-	.31 [.22, .40]
	FP	SS at a distance	0.24 (0.18)	.174	0.10	-	.07 [-.03, .17]
	SP	SS with experience	0.54 (0.13)	< .001	-	1.71	.15 [.08, .21]
	FP	SS with experience	1.44 (0.13)	< .001	-	4.21	.37 [.31, .42]
Bivariate Regression	SP	SS at a distance	0.59 (0.08)	< .001	0.29	-	.29 [.13, .45]
	FP	SS at a distance	0.39 (0.43)	.366	0.04	-	.04 [-.04, .12]
	SP	SS with experience	0.23 (0.08)	.006	-	1.25	.06 [.02, .11]
	FP	SS with experience	2.95 (0.46)	< .001	-	19.08	.63 [.50, .73]
Multiple Regression	SP	SS at a distance	0.62 (0.09)	< .001	0.31	-	.27 [.11, .43]
	FP	SS at a distance	0.74 (0.57)	.193	0.07	-	.06 [-.03, .14]
	SP	SS with experience	0.36 (0.10)	< .001	-	1.44	.10 [.05, .15]
	FP	SS with experience	6.41 (0.72)	< .001	-	605.52	.87 [.81, .91]

Note. SP = summarized preferences, FP = functional preferences, SS = situation selection, OR = odds ratio. Unstandardized regression coefficients (*b*) for situation selection with experience are logit coefficients.

Hypothesis 3 again received support: Summarized preferences for both intelligence and confidence significantly predicted situation selection at a distance across all three approaches. Functional preferences barely predicted situation selection at a distance; effect sizes for both attributes were small and only sporadically significant. Hypothesis 4 received support as well, as functional preferences for both intelligence and confidence predicted situation selection with experience with moderate effect sizes. The effects of summarized preferences on situation selection with experience tended to be much weaker. The double dissociation pattern is most evident in the focal SEM approach, but the pattern with the other two approaches is similar.[10]

Secondary situation-selection dependent measures. To examine whether the findings for H3 could have been driven by incidental differences in the format of our primary dependent measures, we conducted planned analyses on our secondary dependent measures using the same analytic approaches. Specifically, we tested whether summarized preferences would still strongly predict situation selection at a distance if we forced a tradeoff between one website versus another, and whether using a binary choice version of this "at a distance" measure affected the predictive power of functional preferences. Results showed similar patterns of dissociation on the secondary dependent measures, where summarized preferences predicted situation selection at a distance more strongly than functional preferences, regardless of the format of those dependent measures. These results suggest that the support we observed for H3 on the primary dependent measures was not a measurement artifact (see the supplemental materials for details).

Hypotheses 3 and 4: exploring a full model of double dissociation. In addition to the preregistered analyses, we explored the full pattern of double dissociation by fitting a model in

[10] Note that, unlike Study 3, we cannot test the difference between the functional versus summarized preference effect sizes because the two preferences were not entered simultaneously per our analysis plan. We present a test of this idea in the section "Hypotheses 3 and 4: exploring a full model of double dissociation" below. This analysis was preregistered in Study 3 because we conducted Study 3 after we conducted Study 4.

which the primary dependent variables were simultaneously regressed on all attribute preferences we measured (i.e., summarized and functional preferences for both attributes; see Figure 1.8). Each attribute preference predictor was modelled as a latent predictor in the same way as our planned structural equation models. This model allowed us to further isolate the unique predictive validity of each attribute preference variable (e.g., summarized preference for intelligence), controlling for the effects of both the same type of attribute preference for the other attribute (e.g., summarized preference for confidence) and the other type of attribute preference for the same attribute (e.g., functional preference for intelligence).

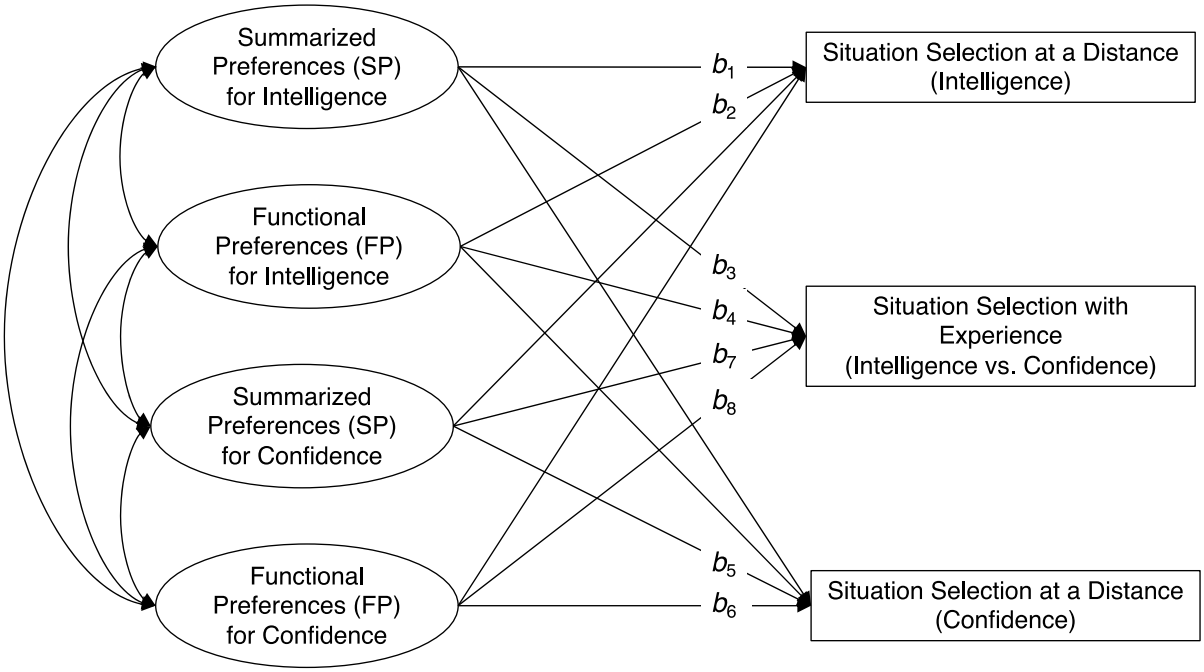

Figure 1.8. Diagram of the full model of double dissociation in which primary dependent variables were regressed on all attribute preferences as latent predictors. Key parameters showing the double dissociation pattern are denoted b_1–b_8 and reported in Tables 1.4A and 1.4B. For visual simplicity, residual (co)variances and measurement model of the latent predictors are not shown.

This model provides a good fit of the data, $\chi^2(83) = 178.16$, $p < .001$, CFI = 0.98, TLI = 0.97, RMSEA = 0.05. Correlations among the attribute preferences are reported in Table 1.3, and

key parameters testing the double dissociation are reported in Tables 1.4A and 1.4B. We observed a full double dissociation between summarized and functional preferences predicting situation selection dependent variables. Summarized preferences predicted situation selection at a distance (H3), and this effect was stronger than the effect for functional preferences (intelligence: $b_1 > b_2$, $\Delta\chi^2(1) = 10.26$, $p = .001$; confidence: $b_5 > b_6$, $\Delta\chi^2(1) = 17.77$, $p < .001$). In contrast, functional preferences predicted situation selection with experience (H4), and this effect was stronger than the effect for summarized preferences (intelligence: $b_4 > b_3$, $\Delta\chi^2(1) = 5.28$, $p = .022$; confidence: $b_8 > b_7$, $\Delta\chi^2(1) = 20.46$, $p < .001$).

Table 1.3.

Correlations Among Latent Predictors in the Full Model of Double Dissociation.

	1	2	3	4
1. Summarized preference for intelligence	-			
2. Functional preference for intelligence	.22***	-		
3. Summarized preference for confidence	.56***	-.02	-	
4. Functional preference for confidence	.18***	.70***	.10*	-

Note. *$p < .05$, **$p < .01$, ***$p < .001$.

Table 1.4A.

Key Parameters for Intelligence in the Full Model of Double Dissociation in Study 4.

Predictor Type	Dependent Variables	Parameter	*b* (*SE*)	*p*	*β* [95%CI]	*OR* [95%CI]
SP	SS at a distance	b_1	0.85 (0.12)	< .001	0.37 [0.28, 0.46]	-
FP	SS at a distance	b_2	0.15 (0.15)	.323	0.06 [-0.06, 0.19]	-
SP	SS with experience	b_3	0.37 (0.14)	.006	-	1.45 [1.11, 1.90]
FP	SS with experience	b_4	0.92 (0.14)	< .001	-	2.51 [1.91, 3.32]

Note. SP = summarized preferences, FP = functional preferences, SS = situation selection, OR = odds ratio. Unstandardized regression coefficients (*b*) for situation selection with experience are logit coefficients.

Table 1.4B.

Key Parameters for Confidence in the Full Model of Double Dissociation in Study 4.

Predictor Type	Dependent Variables	Parameter	*b* (*SE*)	*p*	β [95%CI]	*OR* [95%CI]
SP	SS at a distance	b_5	0.25 (0.13)	< .001	0.39 [0.29, 0.49]	-
FP	SS at a distance	b_6	0.06 (0.15)	.769	0.02 [-0.11, 0.14]	-
SP	SS with experience	b_7	0.35 (0.14)	.011	-	1.42 [1.08, 1.85]
FP	SS with experience	b_8	1.37 (0.13)	< .001	-	3.94 [3.05, 5.08]

Note. SP = summarized preferences, FP = functional preferences, SS = situation selection, OR = odds ratio. Unstandardized regression coefficients (*b*) for situation selection with experience are logit coefficients.

Discussion

The results of Study 4 suggested that although summarized and functional preferences may be only weakly related to each other, both have predictive power. First, we observed correlations between summarized and functional preferences for familiar attributes (i.e., intelligence and confidence) at levels comparable to those of preferences for Reditry in Studies 1–3. Second, we observed a double dissociation between summarized and functional preferences, such that summarized preferences strongly predicted situation selection at a distance (as when people read a description of a website; H3), but functional preferences did so only weakly. In contrast, functional preferences strongly predicted situation selection with experience (as when people see photographs of other website users; H4), but summarized preferences did so only weakly. These results emerged across both of our focal attributes and were similar across a variety of analytic approaches (including when measurement error was taken into account with SEM), thus increasing our confidence in their robustness and generalizability. Moreover, the results from our secondary analyses did not support the possibility that the results for Hypothesis

3 was driven by the particular features of the format of our primary dependent measures: Summarized preferences strongly predicted situation selection at a distance (H3) regardless of whether that dependent variable was measured as interest in a single website (measure i), interest in one website versus the other (measure iii), or a choice between two websites (measure iv). This robust pattern provided strong support for our *a priori* theoretical prediction that people would rely on their summarized preferences to make decisions about situations that they have not yet entered or sampled (see Ledgerwood et al., 2018, Model 3).

General Discussion

People can summarize their attribute preferences (e.g., "I like intelligence in a partner;" "I value loyalty in a friend") and communicate these preferences to others. But where do these summarized preferences come from, and what do they predict? In this research, we set out to investigate the possibility that summarized preferences have some unique antecedents and consequences that distinguish them from functional preferences (e.g., the extent to which intelligence predicts positivity toward a romantic partner).

First, across multiple attributes, summarized and functional preferences correlated at approximately r = .10–.20 (H1). This pattern emerged both in novel learning contexts (where participants reported their summarized preference for a novel trait called Reditry) and familiar contexts (where participants reported their summarized preferences for intelligence and confidence). These correlations may seem modest, given that many literatures have treated these two constructs as interchangeable (e.g., Gerlach et al., 2019). However, these correlations fit comfortably within the range of r = .02–.38 observed for a wide variety of traits in previous large-scale studies of preferences for traits in faces (e.g., Brumbaugh & Wood, 2013). The

current results are consistent with the notion that individuals base their summarized preferences in part on their functional preferences, though perhaps only modestly so.

Second, summarized preference formation was sensitive to incidental features in the learning context that were independent of functional preferences. Specifically, summarized preferences were biased by the likeability of a pool of encountered targets (H2). When participants were asked to form summarized preferences for a novel attribute, they reported that they liked Reditry more when the pool of faces that they encountered during the learning task was more (vs. less) likeable, functional preferences notwithstanding. This effect parallels the outcome density bias in the covariation detection literature: People think a predictor (in this case, Reditry) is more important when the outcome to be predicted (in this case, liking) is common rather than rare. These findings complement earlier work suggesting that another covariation detection bias—the cue-density bias—also affects the formation of summarized preferences in a mating context (Eastwick et al., 2019). Together, these studies suggest that summarized preferences are informed not only by a person's underlying functional preferences for an attribute, but also by other independent features of the learning context.

Lastly, we examined the downstream consequences of summarized and functional preferences. We found that summarized preferences predicted situation selection at a distance, such as the extent to which participants wanted to join a website featuring partners high in Reditry, high in intelligence, or high in confidence (H3). Intriguingly, functional preferences did not predict this outcome especially well. Instead, functional preferences predicted situation selection with experience (i.e., participants' website selection when they saw example profiles of partners high in Reditry, high in intelligence, or high in confidence; H4). Just as with H1, we found evidence of this double dissociation for both a novel attribute in a set of well controlled,

standardized photographs (Study 3), as well as for two familiar attributes in a set of externally valid, naturalistic photographs (Study 4). This pattern of results is consistent with the notion that summarized versus functional preferences (rather than preferences based on appearance vs. extensive experience) differentially predict our situation selection DVs. In other words, by assessing preferences for a novel trait, we eliminated the potential mismatch of evaluative bases of summarized versus functional preferences of known attributes and found further support for the dissociative predictive validity of summarized versus functional preferences. Taken together, the unique antecedents and consequences of summarized preferences lend support to the proposal that summarized and functional preferences are distinct types of evaluative constructs that may serve different psychological purposes.

Implications for Understanding Human Evaluation

Different traditions in the study of attribute preferences. Researchers across multiple disciplines are interested in understanding how humans evaluate attributes. For example, large literatures in the fields of family studies, evolutionary psychology, and close relationships have investigated people's summarized preferences for attributes in a romantic partner (e.g., Buss, 1989; Christensen, 1947; Fletcher et al., 1999; Hill, 1945). Likewise, researchers have examined summarized preferences for attributes of friends, leaders, and teachers (Delaney et al., 2010; Goodwin & Tang, 1991; Pew Research Center Survey, 2015). Meanwhile, researchers who study consumer preferences assess functional preferences for attributes in products (e.g., Delgado & Guinard, 2011; Silayoi & Speece, 2007), researchers who study organizational behavior examine functional preferences for attributes of job candidates and organizations (Heilman & Saruwatari, 1979; Turban & Keon, 1993), and political scientists investigate functional preferences for attributes of election candidates (Carnes & Lupu, 2016).

Across these literatures, researchers tend to assess either summarized preferences or functional preferences following the prevailing measurement tradition in their field. Yet our studies suggest that the distinction between summarized and functional preferences is deeper than a trivial difference in measurement traditions, and researchers should think carefully about which construct they are actually interested in understanding conceptually, and/or what type of outcome they are trying to predict (see also the discussion of measurement correspondence below). Specifically, summarized preferences might be particularly useful when researchers are interested in what people think they like, or contexts in which ideas of liking can be consequential, such as when people introspect about their liking ("I love spiciness in curries!"), and when people communicate their liking with each other ("I can't stand bossiness in a date"). In contrast, functional preferences might be particularly useful when researchers are interested in people's in-the-moment experience of liking, or contexts in which experiences of liking are consequential. When it comes to prediction, researchers may wish to prioritize the assessment of summarized preferences when their goal is to predict decisions at a distance (e.g., whether to visit a destination based on a description in a guidebook; whether to date someone based on an online dating profile). In contrast, researchers may wish to prioritize the assessment of functional preferences when their goal is to predict decisions made with direct experience (e.g., whether to visit a destination for the second time; whether to date someone after meeting them in person; see also Eastwick et al., 2011; Huang et al., 2020).

The large literature on human mate preferences is an interesting case in point. Conceptually speaking, functional preferences are the mate preferences that would have had clearer relevance to ancestral humans; that is, natural selection should have shaped the human mind to positively evaluate real-life mates depending on the extent to which those mates possess

certain attributes (Conroy-Beam et al., 2016). Yet summarized preferences—people's *ideas* about the attributes that appeal to them—are studied far more commonly than functional preferences in the human mate preferences literature (e.g., Buss, 1989; Fletcher et al., 1999), and authors routinely use the word "preference" interchangeably to describe both functional and summarized preferences (e.g., Gerlach et al., 2019; cf. Eastwick et al., 2019). The current findings suggest that researchers studying human mate preferences should make a careful and deliberate choice for any given study about whether to assess functional preferences, summarized preferences, or both. For example, if researchers intend to study a mate selection process that could conceivably be a facsimile of an ancestral selection process, functional preferences are likely the appropriate choice. But researchers also might wish to study the (perhaps uniquely) human ability to draw upon abstract ideas about preferences to guide decisions at a distance (e.g., which outgroup members to meet, which families are suitable for arranging marriages); in these cases, summarized preferences might be especially likely to inform such decisions.

Ideas about liking versus experiences of liking. More broadly, it may be useful to distinguish between people's abstract *ideas* about liking and their concrete *experiences* of liking. In this paper, we have considered this distinction as it applies to attribute preferences, but a similar distinction may be fruitfully applied to attitudes toward objects (i.e., liking for a person, place, or thing; see Ledgerwood et al., 2020). For example, people can have abstract ideas about their liking for broad social categories (e.g., "I like college students") as well as concrete evaluations of specific encountered exemplars (e.g., "I like this particular college student"). Drawing a parallel to the present findings generates the prediction that abstract evaluations of categories would better predict situation selection at a distance (e.g., whether to take a job

described as involving interactions with college students), whereas concrete evaluations of exemplars would better predict situation selection with experience (e.g., whether to take a job after meeting some specific college students in person). A similar distinction exists in the study of attitudinal properties, which differentiates between people's ideas about the affective versus cognitive cause of their attitudes and the actual affective versus cognitive cause of their attitudes (See et al., 2008, 2011). Our work generates the prediction that beliefs about attitudinal properties will have greater relevance to situation selection at a distance, whereas actual attitudinal structure will have greater relevance to situation selection with experience.

The current findings also suggest intriguing hypotheses regarding the consequences of preference-guided situation selection for future research to investigate. To the extent that ideas and experiences of liking diverge, people might select into situations at a distance based on their ideas about liking, but not actually experience more liking once they are in the selected situation (vs. alternative situations). This phenomenon would have implications for myriad real-life contexts. From exclusive dating websites to buying a house, people frequently select themselves into situations and limit the sets of stimuli they subsequently experience based on advertisements, reviews, conversations, and other socially acquired knowledge. People may go to great lengths to enter a certain situation, raise their expectations accordingly, but then not feel as much liking as they anticipated once they have the experience. Future research should examine the possibility that discrepancies between people's ideas about their likes and their actual experiences of liking could create a "cycle of disappointment" along these lines.

Expanding our understanding of measurement correspondence. The present work follows the footsteps of Ajzen and Fishbein's (1977, 2005) classic work on the compatibility principle, which suggests that an attitude will better predict a behavioral criterion when the two

measures correspond in terms of their generality or specificity. A recent resurgence of attention to the importance of considering measurement correspondence has led to new insights and predictions for the study of social influence and implicit bias, as well as attribute preferences (Gawronski, 2019; Irving & Smith, 2020; Ledgerwood et al., 2018; Ledgerwood & Trope, 2010). In a similar vein, our current findings highlight the importance of considering correspondence between measures of attribute preferences and measures of downstream consequences like situation selection.

At the same time, it is important to consider the ways in which the present research expands beyond Ajzen and Fishbein's classic work. Notably, Ajzen and Fishbein (1977, 2005) did not consider the compatibility principle in the context of attitudes toward attributes (see Ledgerwood et al., 2018, for a full discussion). Indeed, the closest analog to the summarized/functional distinction in their work was a distinction between two different measures of *general* attitudes that they treated as interchangeable (Ajzen & Fishbein, 1977): namely, (1) an overall evaluation of a general attitude object (e.g., a person's favorability toward maintaining good physical health) and (2) the average of a series of evaluations of specific attitude objects (e.g., a person's average favorability toward eating more vegetables, avoiding junk food, exercising daily, and getting regular checkups). Because summarized and functional preferences would have been treated as two forms of general attitudes, a prediction from Ajzen and Fishbein's conceptualization would be that summarized and functional preferences should predict outcomes equally. In contrast, we posit that summarized and functional preferences are distinct: Summarized preferences are abstract evaluations of attributes as concepts, whereas functional preferences are concrete evaluations of attributes as experiences. Drawing from work on construal level fit (Fujita et al., 2008; Lee et al., 2010), we predicted and found summarized

preferences strongly predicted situation selection at a distance, whereas functional preferences strongly predicted situation selection with experience. These findings, as part of a double dissociation, is (broadly speaking) a form of correspondence, but it is not derivable from the Fishbein and Ajzen correspondence principle, which predated construal level theory.

Perhaps most importantly, the existence of divergent measurement traditions for assessing attribute preferences suggests that the issue of correspondence has yet to receive sufficient attention in these literatures. By demonstrating the distinct predictive validity of summarized and functional attribute preferences, our work highlights the importance of considering measurement compatibility for future research on attribute preferences in human mating, consumer preferences, organizational behavior, and beyond.

Strengths and Limitations

Drawing from multiple literatures on attribute preferences, we replicated prior work on the relations between summarized and functional preferences for familiar attributes and we tested novel hypotheses on the distinct antecedents and consequences of summarized preferences. We tested these new hypotheses using well-powered studies, preregistered analyses, and both experimental and correlational designs, which together give us confidence that our results are likely robust. In addition, the online dating context used in Study 4 had the advantage of mimicking real-life situation selection and allowing us to manipulate how the situations were presented.

However, further research is needed to test the extent to which our findings will generalize beyond contexts in which people evaluate photographs and participate in online dating. On the one hand, consider that summarized and functional preferences of attributes correlate more strongly in less complex, nonsocial objects (e.g., juices; Alcser et al., 2021).

Therefore, the double dissociation in downstream predictive consequences might weaken or disappear when people select into situations involving nonsocial objects. On the other hand, summarized and functional preferences are effectively uncorrelated for attributes perceived via naturalistic face-to-face interactions (Ledgerwood et al., 2018; Eastwick et al., 2021; Sparks et al., 2020). It would be useful for future research to examine these contexts, too.

Future research can also further clarify the relation between functional and summarized preferences. For example, functional-summarized preference correlations may be relatively weak because functional preferences are not accessible and/or not diagnostic when people report their summarized preferences (Feldman & Lynch, 1988). It might be possible to gather evidence for these mechanisms by incorporating existing manipulations of accessibility (e.g., filler tasks; Ahluwalia & Gurhan-Canli, 2000) or diagnosticity (e.g., instructions that using a certain information is "good"; Zhang & Khare, 2009) and observing whether the functional–summarized preference correlation shifts accordingly.

Finally, although other studies have manipulated functional preferences to affect summarized preferences (Eastwick et al., 2019, Study 1), the current studies only measured functional and summarized preferences. Therefore, it is certainly possible that summarized preferences affected functional preferences in addition to the reverse pathway we depict in Figure 1.1, perhaps especially with the familiar attributes in Study 4.

Conclusions

The current research provides an important first step in understanding the predictive power of summarized and functional preferences and begins to delineate when and how summarized preferences may be useful. Going forward, we believe the interdisciplinary study of

attribute preferences will greatly benefit from researchers carefully considering which preference construct they are interested in understanding and which outcomes they are trying to predict.

Supplemental Material

Study S1

In this study, we tested whether experimentally manipulating a set of target faces to be more (vs. less) likeable would lead participants to infer stronger summarized preferences for a novel attribute. We also assessed the strength of the association between summarized and functional preferences for this attribute. The procedure was similar to Studies 1 and 2, except that we used both male faces and female faces in this study. Participants evaluated either female faces or male faces based on the sex they were primarily attracted to.

Method

Female faces.

Participants and power. One hundred and one participants primarily attracted to females completed the study online through Amazon's Mturk platform. They were randomly assigned to one of two between-subjects conditions (low average likeability vs. high average likeability). We decided *a priori* to target a cell size of 50 participants per cell based on our lab's standard practice for minimum cell size (the total number of completed surveys in Qualtrics ended up being slightly higher).

We set and recorded the following *a priori* exclusion criteria: We would exclude participants who (1) gave an identical rating to all faces presented for measurement of functional preferences, and/or (2) provided a nonsensical response to a Winograd-like schema designed to filter out bots or inattentive participants. The number of participants who met each of these exclusion criteria were $n = 0$ and $n = 3$, respectively, resulting in a final sample of $N = 98$ (11 females, 85 males, and 2 people who chose another option; $M_{age} = 33.7$, $SD = 9.7$).

Procedure. The procedure was identical to Study 1. First, participants saw a series of 24 faces, each presented along with its level of Reditry, and rated their romantic liking for each pictured person. After the trials, participants completed a measure of their overall summarized preference for Reditry. Lastly, after seeing another survey unrelated to the current research questions, participants completed the attention check and a demographic survey.

New materials and measures.

Stimuli. We selected 48 White female faces from the Chicago Faces Database (CFD; Ma, Correll, & Wittenbrink, 2015). To manipulate average likeability, we divided the faces into two sets of 24 faces that varied similarly in babyfacedness (according to the norming-data ratings in Ma et al., 2015) and that differed only in how likeable they were on average. In a previously published sample (Eastwick & Smith, 2018), $N = 677$ participants who were primarily attracted to women evaluated each face on a measure of romantic likeability using 1-7 rating scales. The faces we chose for the high likeability condition had a mean of $M = 3.49$ ($SD = 0.89$) on this scale, and the faces we chose for the low likeability condition had a mean of $M = 2.10$ ($SD = 0.59$). To avoid unintentionally manipulating the strength of the association between babyfacedness and likeability, we ensured that the correlations between the Ma et al. (2015) ratings of babyfacedness and the Eastwick and Smith (2018) ratings of likeability were similar across conditions ($r = .25$ in the high likeability condition and $r = .26$ in the low likeability condition); we also checked that this correlation was similar to the correlation between babyfacedness and likeability in the full population of White female faces in the

CFD (r = .28). Finally, we inspected the scatterplot between these two variables (see Figure S1.1).

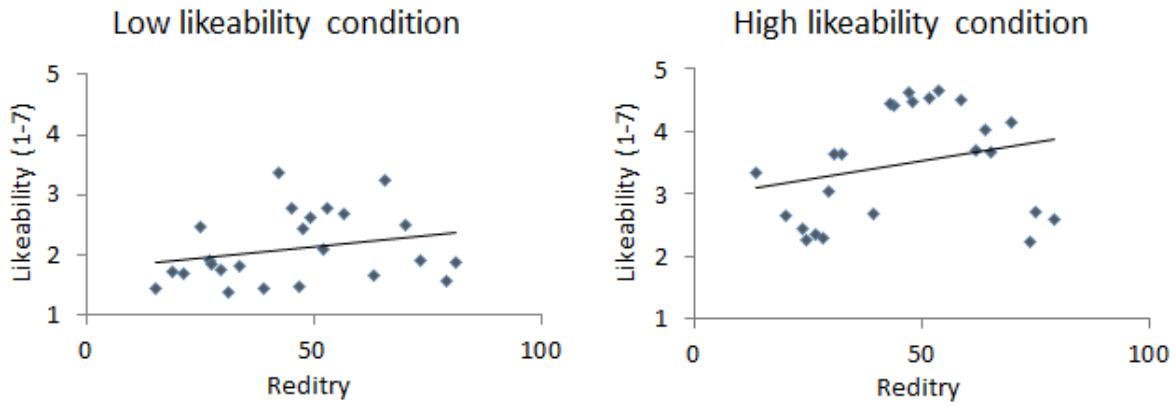

Figure S1.1. Scatterplots of the stimuli; each dot represents a female face target. Notice that the correlation between pretest ratings of likeability and Reditry is the same in both conditions, whereas the average likeability is higher in the high (vs. low) likeability condition.

Male faces.

Participants and power. One hundred and twenty-three participants primarily attracted to males completed the study online through Amazon's Mturk platform. They were randomly assigned to one of two between-subjects conditions (low average likeability vs. high average likeability). The power calculation and exclusion criteria were identical to those described above. The number of participants who gave identical ratings was n = 3 and the number of those who provided a nonsensical response to the Winograd-like schema were n = 3, resulting in a final sample of N = 117 (107 females, 10 males, and 1 person who chose another option; M_{age} = 36.9, SD = 12.8).

Procedure. The procedure was identical to Studies 1 and 2, as well as the one described above for female faces.

New materials and measures.

Stimuli. We used the same male faces as we did in Study 2. As a reminder, the average likeability of faces in the high likeability condition was $M = 2.76$ ($SD = 0.59$), and the average likeability of faces in the low likeability condition was $M = 1.83$ ($SD = 0.38$). We ensured that the correlation between pretest ratings of babyfacedness and likeability was similar across conditions ($r = .05$ in both conditions; see Figure S1.2) and reflected the actual correlation in the full population of White male faces in the Chicago Faces Database ($r = .01$).

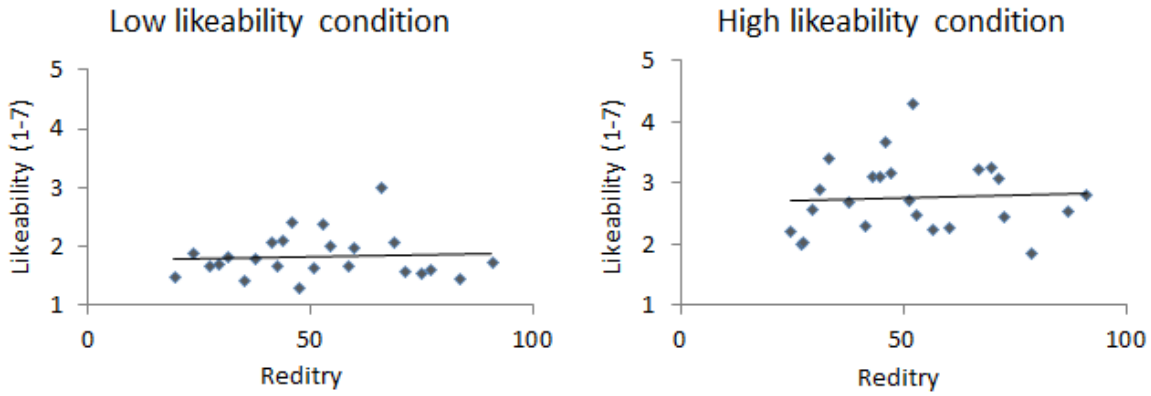

Figure S1.2. Scatterplots of the stimuli; each dot represents a male face target. Notice that the correlation between pretest ratings of likeability and Reditry is the same in both conditions, whereas the average likeability is higher in the high (vs. low) likeability condition.

Results

Female faces.

Manipulation check. We checked whether the manipulation of average target likeability successfully influenced the amount of liking that participants experienced when learning about their preferences. Our manipulation of average target likeability was successful: On average, participants in the high likeability condition experienced greater liking for the faces they saw ($M = -0.22$, $SD = 1.47$) than participants in the low

likeability condition (M = -1.65, SD = 1.33), $t(96) = 5.04$, $p < .001$, $d = 1.02$, 95% CI [0.60, 1.44].

Functional preferences for Reditry. Although we took care to ensure that the correlation between Reditry and pretest ratings of face likeability were similar across conditions, our manipulation of average likeability unintentionally influenced participants' experienced functional preferences for Reditry, such that their functional preferences for Reditry was higher in the high (vs. low) likeability condition (M = 0.24, SD = 0.30 vs. M = 0.13, SD = 0.16), $t(96) = 2.32$, $p = .022$, $d = 0.46$, 95% CI [0.06, 0.86]. This result underscores the difficulty of selecting appropriate stimuli and the importance of testing different sets of stimuli.

Main analyses. We tested whether functional and summarized preferences were correlated (H1). The correlation between functional and summarized preferences was $r = .06$, $p = .532$, 95% CI [-.14, .26].

We planned *a priori* to test the effect of the likeability manipulation on summarized preferences (H2). The results showed that summarized preferences were higher in the high vs. low likeability condition, M = 0.20, SD = 1.68 vs. M = -0.98, SD = 2.27, $t(96) = 2.94$, $p = .004$, $d = 0.59$, 95% CI [0.19, 0.97]. Because our manipulation unintentionally affected functional preferences (see above), our interpretation of this effect is ambiguous: The difference in functional preferences could be driving this effect, instead of the hypothesized difference in likeability. Although this alternative interpretation is unlikely, given that summarized and functional preferences were uncorrelated in this sample, we are hesitant to draw conclusions from this result.

Male faces.

Manipulation check. We checked whether the manipulation of average target likeability successfully influenced the amount of liking that participants experienced when learning about their preferences. Our manipulation of average target likeability was successful: On average, participants in the high likeability condition experienced greater liking for the faces they saw (M = -1.07, SD = 1.35) than participants in the low likeability condition (M = -1.93, SD = 1.22), $t(115) = 3.53$, $p < .001$, d = 0.67, 95% CI [0.29, 1.05].

Functional preferences for Reditry. Functional preferences were very similar across the two conditions (M = 0.03, SD = 0.31 vs. M = 0.00, SD = 0.23), $t(115) = 0.61$, p = .543, d = 0.12, 95% CI [-0.49, 0.25], confirming that our manipulation of average target likeability did not affect participants' functional preferences for Reditry.

Main analyses. After confirming that our manipulation was successful at influencing average liking but not functional preferences for Reditry, we proceeded to our main analyses. First, we tested whether functional and summarized preferences were correlated (H1). The correlation between functional and summarized preferences was r = .17, p = .067, 95% CI [-.01, .34].

Next, we tested whether average likeability of the targets biased summarized preferences for Reditry (H2). Indeed, participants inferred stronger summarized preferences for Reditry in the high versus low likeability conditions (M = -0.58, SD = 1.94 vs. M = -1.34, SD = 1.90), $t(115) = 2.11$, p = .037, d = 0.40, 95% CI [0.03, 0.77]. In other words, participants inferred that they liked Reditry substantially more when they learned about their preference in a context with high (vs. low) likeability targets.

Study S2

Throughout this manuscript, we use the conceptual term "romantic liking" interchangeably with the terms "romantic interest" and "romantic desire;" these latter two terms were used in the actual items that participants rated in the studies. But are these terms sufficiently highly associated that it is appropriate to treat them synonymously? To test this idea, we asked participants to rate the faces used in Studies 1 and 2 on these three items.

Method

Participants. One hundred and nine participants completed the study online through Amazon's Mturk platform. We decided *a priori* to collect data from at least 100 participants. Consistent with all other studies, we limited participants to those who were between 18-35 years old. We set and recorded the following *a priori* exclusion criteria: We would exclude participants who (1) gave an identical rating to all faces presented, (2) provided a nonsensical response to a Winograd-like schema designed to filter out bots or inattentive participants, and/or (3) expressed suspicion about the purpose of the study. One participant met the second exclusion criterion and was excluded, resulting in a final sample of N = 108 (56 females, 51 males, and 1 person who chose another option; M_{age} = 28.1, SD = 4.0).

Procedure. Participants first completed a brief prescreen in which they indicated their age and the sex to which they were primarily romantically attracted, which determined the sex of the potential partners presented to them throughout the rest of the study. Participants then saw a series of 48 faces; the female faces were the same as those presented in Study 1, and the male faces were the same as those presented in Study 2. Participants rated each picture person on romantic interest ("To what extent are you romantically interested in this person?" -4 = *strongly dislike*, -4 = *strongly like*), romantic liking ("To what extent do you romantically like this person?" -4 = *strongly dislike*, 4 = *strongly like*), and romantic desire ("To what extent do you

experience romantic desire for this person?" 1 = *not at all*, 9 = *a great deal*), all on 9-point Likert-type scales. The question on romantic interest was worded to exactly match that of the measure of functional preferences in Studies 1 and 2, and the question on romantic desire was worded to exactly match that of the measure of functional preferences in Study 3. Lastly, participants completed the same attention check as used in Studies 1 and 2, a demographic survey, and a short questionnaire unrelated to the current research questions.

Results and Discussion

As primary analyses, we conducted cross-classified mixed effects modeling to test the relation between ratings on each pair of variables (romantic interest, romantic liking, and romantic desire). Because each participant rated each variable across faces, cross-classified mixed effects models allowed us to distinguish between the fixed effects among the variables (e.g., the relation between interest and liking) and the random effects of stimuli and participants (Raudenbush & Bryk, 2002). Accounting for the nested nature of the data at both the participant and the stimulus level also provided us with more accurate estimates (Judd, Westfall, & Kenny, 2012, 2017). We ran three models: one in which desire was regressed on interest, one in which interest was regressed on liking, and one in which liking was regressed on desire. Our models were specified as follows:

$$\text{Level 1: } Y_{ijk} = \gamma_{0jk} + \gamma_{1jk}X_{ijk} + e_{ijk}$$

$$\text{Level 2: } \gamma_{0jk} = \gamma_{000} + b_{00j} + c_{00k} + d_{0jk}$$

At the first level, Y_{ijk} is the rating on the dependent variable (e.g., interest) on the *i*th trial by participant *j* responding to the *k*th face, and X_{ijk} is the rating on the independent variable (e.g., liking) on the *i*th trial by participant *j* responding to the *k*th face. The intercept γ_{0jk} is the mean rating by participant *j* to the *k*th face, the coefficient γ_{1jk} is the fixed effect of the

independent variable, and the error term e_{ijk} is the residual of the model, $e_{ijk} \sim N(0, \sigma^2)$. At the second level, γ_{000} is the overall mean rating across all trials, b_{00j} is the random main effect of participant *j* (averaged over all faces), c_{00k} is the random main effect of the *k*th face (averaged over all participants), and d_{0jk} is the random interaction effect of participant by face. Ratings on each variable were standardized across faces and participants (i.e., grand-mean centered).

Results showed that the three variables were highly associated with each other. Romantic interest significantly predicted romantic desire, β = .87, 95% CI [.86, .89], romantic liking significantly predicted romantic interest, β = .93, 95% CI [.92, .94], and romantic desire significantly predicted romantic liking, β = .84, 95% CI [.82, .86], all *p*s < .001. We also conducted secondary analyses in which we calculated between-subjects correlations among the three variables by averaging ratings by each participant across faces, and the results were highly similar: $r_{\text{desire.interest}}$ = .88, 95% CI [.82, .92], $r_{\text{interest.liking}}$ = .99, 95% CI [.98, .99], $r_{\text{liking.desire}}$ = .86, 95% CI [.80, .91]. Therefore, we concluded that participants' ratings on romantic interest, romantic liking, and romantic desire were interchangeable.

Study S3

In Study 4, we used photographs that we collected for use as stimuli from a publicly accessible dating website. Following the stimuli collection procedure of Wood and Brumbargh (2009), we set three criteria for the selected photographs, which were required to (1) show at least the person's head and torso in full view, (2) be of a reasonably high quality (i.e., not blurry or unfocused, or so small that facial features cannot be discerned), and (3) contain only one individual. Consistent with Wood and Brumbargh (2009), we selected all photographs from the "aged 18 to 25" range on the website and stopped stimuli collection once 100 photographs per target sex met our criteria to ensure a random cross-section of photographs.

Following stimuli collection, we conducted a separate rating study. Participants (N = 132; 71 women, 61 men) between the ages of 18 and 35 years (M = 28.83, SD = 4.04) completed the study online through MTurk. Participants rated each of the 100 preferred-sex targets on a list of attributes that people can rapidly and consensually rate on from faces (Oosterhof & Todorov, 2008). On each screen, participants saw one target and the list of attributes below the target's photograph, and participants rated the target on each attribute on a 9-point scale (1 = *not at all*, 9 = *extremely*). Cronbach's alphas were high for ratings on both "intelligent" (α = .95 for male targets and α = .86 for female targets) and "confident" (α = .94 for male targets and α = .88 for female targets; see Table S1.1 for the full list of attributes rated and their descriptive statistics). Table S1.1.

*Internal Consistency (Cronbach's Alpha) and Interrater Agreement (*r*) of Ratings on Attributes*

Attribute	Cronbach's alpha (α)		Interrater agreement (*r*)	
	Male Targets	Female Targets	Male Targets	Female Targets
Attractive	.96	.97	.27	.28
Mean	.93	.80	.17	.07
Dominant	.94	.84	.18	.08
Trustworthy	.94	.80	.17	.06
Aggressive	.94	.81	.18	.06
Caring	.93	.78	.18	.06
Emotionally stable	.93	.84	.18	.07
Responsible	.95	.85	.21	.08
Sociable	.93	.86	.18	.09
Confident	.94	.88	.21	.10
Intelligent	.95	.86	.21	.08
Sensitive	.93	.73	.16	.04

Note: N = 66 for ratings of 100 male targets, and N = 66 for ratings of 100 female targets.

Study 4: Additional Results

Hypotheses 3 and 4: Preregistered Analyses

Primary situation-selection dependent measures.

Table S1.2.

Fit Indices from Structural Equation Models with Summarized and Functional Preferences Predicting Primary Dependent Variables in Study 4.

Predictor Type	Dependent Variables	Attributes	χ^2	*df*	*p*	CFI	TLI	RMSEA
SP	SS_d	Intelligence	10.34	7	.170	1.00	1.00	0.03
		Confidence	8.96	7	.256	1.00	1.00	0.02
FP	SS_d	Intelligence	265.30	21	< .001	0.90	0.84	0.14
		Confidence	263.68	21	< .001	0.90	0.84	0.14
SP	SS_e	Intelligence & Confidence	1.23	7	.990	1.00	1.01	0.00
FP	SS_e	Intelligence & Confidence	125.86	21	< .001	0.96	0.94	0.09

Note: SP = summarized preferences, FP = functional preferences, SS_d = situation selection at a distance, SS_e = situation selection with experience. In all analyses, preferences for both intelligence and confidence were entered as predictors. Because situation selection with experience was a dichotomous choice between two situations (i.e., website with highly intelligent targets and website with highly confident targets), the summarized preference model and the functional preference model estimated the effects of preferences for intelligence and confidence on this DV simultaneously. Therefore, we report only one set of fit indices for each predictor type on that variable. For analyses involving the dichotomous dependent variable (i.e., situation selection with experience), fit indices were calculated using the diagonally weighted least squares (DWLS) estimator.

Secondary situation-selection dependent measures. To examine whether the observed double dissociation between summarized and functional preferences could have been driven by an incidental feature of the format of our primary dependent measures, we conducted planned analyses on our secondary dependent measures using the same analytic approaches. First, we asked whether summarized preferences would still strongly predict situation selection at a distance if we forced a tradeoff between one website versus another. After all, the double dissociation observed in our primary analyses could be driven by a difference in how people responded to a single situation (as measured by the primary situation selection at a distance variable above) rather than a tradeoff between two situations (as measured by the primary

situation selection with experience variable above). We examined this possibility by assessing how strongly summarized and functional preferences predicted participants' interest in one website *versus* the other on a bipolar rating scale (i.e., the website that would provide access to highly intelligent partners versus the website that would provide access to highly confident partners; see Table S1.3 for a summary of effect sizes, and Table S1.4 for the relevant fit indices; all models fit the data well).

Table S1.3.

Effect Sizes for Summarized and Functional Preferences Predicting Secondary Dependent Variables in Study 4.

Analytic Approaches	Predictor Type	Dependent Variables	Attributes	
			Intelligence	Confidence
Structural Equation Models	SP	SS at a distance (tradeoff)	.48***	.38***
	FP	SS at a distance (tradeoff)	.20***	.13**
	SP	SS at a distance (choice)	.28***	.26***
	FP	SS at a distance (choice)	.12**	.12**
Bivariate Regression	SP	SS at a distance (tradeoff)	.31***	.11*
	FP	SS at a distance (tradeoff)	.15***	-.04
	SP	SS at a distance (choice)	.12***	.09***
	FP	SS at a distance (choice)	.18	.13
Multiple Regression	SP	SS at a distance (tradeoff)	.41***	.29***
	FP	SS at a distance (tradeoff)	.16***	.08
	SP	SS at a distance (choice)	.25***	.23***
	FP	SS at a distance (choice)	.45**	.39*

Note: SP = summarized preferences, FP = functional preferences, SS = situation selection. *p < .05, **p < .01, ***p < .001. All effect sizes are reported in correlation coefficients.

Results from our focal SEM approach revealed that summarized preferences strongly predicted this tradeoff version of the measure (intelligence: b = 1.27, SE = 0.12, p < .001, r = .48, 95% CI [.40, .57]; confidence: b = 0.98, SE = 0.12, p < .001, r = .38, 95% CI [.29, .46]). That is, when we equated our situation selection at a distance and situation selection with experience measures in terms of both forcing a tradeoff, we still see that summarized preferences strongly predicted situation selection at a distance (Table S1.3), whereas summarized preferences weakly

predicted situation selection with experience (see Table 1.2). The results were similar across the two traits, and mostly similar across the two alternative analytic approaches.

Table S1.4.

Fit Indices from Structural Equation Models with Summarized and Functional Preferences Predicting Secondary Dependent Variables in Study 4.

Predictor Type	Dependent Variables	Attributes	χ^2	*df*	*p*	CFI	TFI	RMSEA
SP	SS_d (tradeoff)	Intelligence & Confidence	7.81	7	.350	1.00	1.00	0.01
	SS_d (choice)	Intelligence & Confidence	2.55	7	.923	1.00	1.01	0.00
FP	SS_e (tradeoff)	Intelligence & Confidence	264.36	21	< .001	0.91	0.84	0.14
	SS_e (choice)	Intelligence & Confidence	53.89	21	< .001	0.99	0.98	0.05

Note: SP = summarized preferences, FP = functional preferences, SS_d = situation selection at a distance, SS_e = situation selection with experience. Because situation selection with experience was a dichotomous choice between two situations (i.e., website with highly intelligent targets and website with highly confident targets), one model simultaneously estimated the effects of preferences for both attributes for each predictor type. Therefore, we report only one set of fit indices for each predictor type on that variable. For all analyses, fit indices were calculated using the DWLS estimator.

We also explored whether forcing a tradeoff affected the predictive power of functional preferences. The effects of functional preferences on the tradeoff measure were more ambiguous (Table S1.3). Across the different attributes and analytic approaches, some of the effect sizes on the tradeoff measure were more similar to those of functional preferences on our original measure of situation selection at a distance, and some of the effect sizes on the tradeoff measure were more similar to those of functional preferences on situation selection with experience (see Table 1.2). These intermediate results suggest that there may be something special about tradeoffs that gives functional preferences a little extra predictive power. However, because the results from functional preferences predicting the tradeoff measure were less consistent across the two attributes and alternative analytic approaches, it would be important to replicate these

results and find a more consistent pattern across analyses and attributes before drawing any strong conclusions from them.

Next, we asked whether summarized preferences would still strongly predict participants' situation selection at a distance if we asked them to make a binary choice between one website versus another. After all, the double dissociation observed in our primary analyses could be driven by a difference between using a rating scale to evaluate situation desirability (as measured by our primary situation selection at a distance variable) and making a binary choice between situations (as measured by our primary situation selection with experience variable). We examined this possibility by assessing how strongly summarized and functional preferences predicted participants' choice between the two described websites. Results from our focal SEM approach revealed that summarized preferences still strongly predicted this choice version of the "at a distance" measure (intelligence: $b = 1.04$, $SE = 0.10$, $p < .001$, $r = .28$, 95% CI [.23, .32]; confidence: $b = 1.00$, $SE = 0.13$, $p < .001$, $r = .26$, 95% CI [.20, .33]). That is, when we equated our situation selection at a distance and situation selection with experience measures in terms of both involving a binary choice, we still saw that summarized preferences strongly predicted situation selection at a distance (Table S1.3), whereas recall that summarized preferences weakly predicted situation selection with experience (see Table 1.2). The results were similar across the two attributes and across the alternative analytic approaches.

We can also explore whether using a binary choice version of this "at a distance" measure affected the predictive power of functional preferences. The results from our focal SEM approach revealed that functional preferences weakly predicted the binary version of situation selection at a distance (intelligence: $b = 0.44$, $SE = 0.16$, $p = .005$, $r = .12$, 95% CI [.04, .20]; confidence: $b = 0.45$, $SE = 0.16$, $p = .004$, $r = .12$, 95% CI [.04, .21]). In other words, when we

equate our situation selection at a distance and situation selection with experience measures in terms of both involving a binary choice, we still see that functional preferences weakly predicted situation selection at a distance (Table S1.3), whereas recall that functional preferences strongly predicted situation selection with experience (see Table 1.2). The results were similar across the two attributes and mostly similar across the alternative analytic approaches.[11]

Exploratory analyses on retaining the item "charismatic" in summarized preferences for confidence. We explored the impact of retaining the item "charismatic" in our calculation of summarized preference for confidence on predictions by summarized preferences. The results did not change in any substantive way (see Tables S1.5A and S1.5B). In addition, all five structural equation models with summarized preferences as predictors fit the data at a level comparable with their corresponding models reported in the manuscript: summarized preference for intelligence predicting situation selection at a distance, $\chi^2(12) = 21.20$, $p = .047$, CFI = 0.99, TLI = 0.99, RMSEA = 0.04; summarized preference for confidence predicting situation selection at a distance, $\chi^2(12) = 24.41$, $p = .018$, CFI = 0.99, TLI = 0.99, RMSEA = 0.04; summarized preferences predicting situation selection with experience, $\chi^2(12) = 6.11$, $p = .911$, CFI = 1.00, TLI = 1.01, RMSEA = 0.00; summarized preferences predicting situation selection with experience (tradeoff), $\chi^2(12) = 18.74$, $p = .095$, CFI = 1.00, TLI = 0.99, RMSEA = 0.03; summarized preferences predicting situation selection (choice), $\chi^2(12) = 6.76$, $p = .873$, CFI = 1.00, TLI = 1.00, RMSEA = 0.00.

Table S1.5A.

[11] Note that it is complicated to compare the strength of the predictive power of summarized versus functional preferences directly. For example, although the effect sizes for functional preferences predicting situation selection at a distance (choice) from the bivariate regressions and multiple regressions were nominally larger than those for summarized preferences, they were less significant (i.e., their associated *p*-values were larger) due to greater uncertainty around their estimates. In contrast, one can more readily compare the magnitude of the coefficients for summarized preferences predicting different dependent measures, and the magnitude of the coefficients for functional preferences predicting different dependent measures, due to comparable standard errors.

Test Statistics and Effect Sizes for Summarized Preferences for Intelligence Predicting Primary and Secondary Dependent Variables, with Three Indicators of Summarized Preference for Confidence.

Analytic Approaches	Dependent Variables	*b* (*SE*)	*p*	*β*	*OR*	*r*
Structural	SS at a distance	0.89 (.13)	< .001	.39	-	.32
Equation	SS with experience	0.45 (.13)	< .001	-	1.56	.12
Models	SS at a distance (tradeoff)	1.30 (.12)	< .001	.60	-	.49
	SS at a distance (choice)	1.10 (.10)	< .001	-	3.02	.29
Bivariate	SS at a distance	0.73 (.09)	< .001	.33	-	.33
Regression	SS with experience	0.13 (.10)	.192	-	1.13	.03
	SS at a distance (tradeoff)	0.66 (.09)	< .001	.31	-	.31
	SS at a distance (choice)	0.44 (.09)	< .001	-	1.55	.12
Multiple	SS at a distance	0.73 (.10)	< .001	.33	-	.29
Regression	SS with experience	0.31 (.11)	.005	-	1.37	.09
	SS at a distance (tradeoff)	1.04 (.09)	< .001	.49	-	.43
	SS at a distance (choice)	0.98 (.14)	< .001	-	2.67	.26

Note: SS = situation selection. Unstandardized regression coefficients (*b*) for dichotomous variables are logit coefficients.

Table S1.5B.

Test Statistics and Effect Sizes for Summarized Preferences for Confidence Predicting Primary and Secondary Dependent Variables, with Three Indicators of Summarized Preference for Confidence.

Analytic Approaches	Dependent Variables	*b* (*SE*)	*p*	*β*	*OR*	*r*
Structural	SS at a distance	0.89 (.13)	< .001	.39	-	.32
Equation	SS with experience	0.52 (.13)	< .001	-	1.69	.14
Models	SS at a distance (tradeoff)	1.02 (.12)	< .001	.47	-	.39
	SS at a distance (choice)	1.05 (.13)	< .001	-	2.85	.28
Bivariate	SS at a distance	0.68 (.09)	< .001	.32	-	.32
Regression	SS with experience	0.20 (.09)	.020	-	1.23	.06
	SS at a distance (tradeoff)	0.24 (.09)	.005	.12	-	.12
	SS at a distance (choice)	0.33 (.10)	.001	-	1.39	.09
Multiple	SS at a distance	0.75 (.10)	< .001	.35	-	.30
Regression	SS with experience	0.36 (.10)	< .001	-	1.43	.10
	SS at a distance (tradeoff)	0.75 (.09)	< .001	.36	-	.32
	SS at a distance (choice)	0.93 (.14)	< .001	-	2.53	.25

Note: SS = situation selection. Unstandardized regression coefficients (*b*) for dichotomous variables are logit coefficients.

Chapter 2

Evaluations of Empathizers Depend on the Target of Empathy

Cite: Wang, Y. A., & Todd, A. R. (in press). Evaluations of empathizers depend on the target of empathy. *Journal of Personality and Social Psychology*.

Abstract

Psychological research on empathy typically focuses on understanding its effects on empathizers and empathic targets. Little is known, however, about the effects of empathy beyond its dyadic context. Taking an extra-dyad perspective, we examined how third-party observers evaluate empathizers. Seven experiments documented that observers' evaluations of empathizers depend on the target of empathy. Empathizers (vs. non-empathizers) of a stressful experience were respected/liked more when the empathic target was positive (e.g., children's hospital worker), but not when the target was negative (e.g., white supremacist; Experiments 1–2). Empathizers were respected/liked more when responding to a positive target who disclosed a positive experience (i.e., a personal accomplishment), but *less* when responding to a negative target who disclosed a positive experience (Experiment 3). These effects were partly, but not solely, attributable to the positivity of empathic responses (Experiment 4). Expressing empathy (vs. condemnation) toward a negative target resulted in *less* respect/liking when the disclosed experience was linked to the source of target valence (i.e., stress from white supremacist job; Experiments 5–7), but *more* respect/liking when the experience was unrelated to the source of target valence (i.e., stress from cancer; Experiment 7). Overall, empathizers were viewed as warmer, but to a lesser extent when responding to a negative target. These findings highlight the importance of considering the extra-dyad impact of empathy and suggest that although people are often encouraged to empathize with disliked others, they are not always favored for doing so.

Keywords: attitudes; empathy; impression formation; perspective taking; person perception

Introduction

In November, 2017, a *New York Times* article by journalist Richard Fausset drew harsh criticism from the public. The article profiled a man named Tony Hovater and depicted mundane details from his life, including the contents of his wedding registry, TV shows he enjoys, and his music preferences. The journalist took an empathic approach to understand why "…this man, intelligent, socially adroit and raised middle class…gravitate[s] toward the furthest extremes of American political discourse" (Fausset, 2017). The profile was derided because Hovater is a white nationalist. "Nazi sympathizers are supposed to be reviled and ostracized, not humanized and normalized," a reader wrote to the editor (Shapiro, 2017). Other readers similarly chastised the journalist for expressing empathy toward Hovater, claiming instead that he should have been more neutral or even actively condemning (e.g., Vernon, 2017).

The backlash to this profile illustrates that expressions of empathy—typically studied at a dyadic level between expressers of empathy (i.e., empathizers) and the recipients of those expressions (i.e., empathic targets)—can have a broader impact on people outside the dyad. Despite the vast literature on empathy and the increasingly central role it plays in public discourse (e.g., Baron-Cohen, 2011; Decety & Ickes, 2009), however, current understanding of empathy largely remains limited to the empathic dyad. The view emerging from this literature is generally positive: Empathy is often celebrated as a moral virtue, and expressions of empathy are evaluated favorably by *targets* (e.g., Goldstein et al., 2014). Understanding how *third-party observers* evaluate empathy, especially how they evaluate empathizers, not only promises to advance theoretical understanding of the social effects of empathy; it also has practical implications for understanding how empathy affects social networks, where observers'

evaluations can have consequences for people in empathic dyads. Here, we examined how third-party observers—those who witness expressions of empathy as outsiders—evaluate empathizers.

Evaluations of Empathy

Empathy is broadly conceptualized as a multifaceted, interpersonal construct (Batson, 2009; Davis, 1994). Although many definitions of empathy exist, most definitions include cognitive and affective components that entail acknowledging (and sometimes sharing) how another person thinks or feels (e.g., Decety & Hodges, 2004; Zaki & Ochsner, 2012). Empathy is commonly viewed as a "universal good." Buddhist traditions consider "empathic joy" a human ideal (Davidson & Harrington, 2002; Wallace & Shapiro, 2006). Philosopher Adam Smith (1759/1976) claimed that the ability to "place ourselves in [another's] situation…and become in some measure the same person with him," is essential to moral good. More recently, empathy has been hailed by politicians, entrepreneurs, and scholars as a key path toward various forms of social flourishing, including justice, intergroup harmony, global peace, and even human survival (e.g., Baron-Cohen, 2011; Obama, 2006; Rifkin, 2009; Safire, 2008). This call for empathy parallels the growing popularity of empathy training in the workplace and the classroom (Crowley & Saide, 2016; Lublin, 2016; Spencer-Keyse, 2018). Empathy, it seems, is a virtue believed to improve social relations and to shape the next generation for the better.

Why is empathy so fervently advocated? One putative benefit of empathy is that it can help bridge social divides. This idea can be traced to various cultural roots: For example, religious teachings explicitly encourage empathy toward people who are different from oneself—even people one may actively dislike. Christians are taught to "love your enemies." This sentiment is echoed in a Sioux prayer: "Great Spirit, help me never to judge another until I have walked in his moccasins." Contemporary perspectives likewise maintain that empathy

across social divides enables prosocial outcomes and that intergroup conflict results, in part, from empathic failures (Klimecki, 2019; Todd & Galinsky, 2014; Zaki & Cikara, 2015). For example, the Center for Empathy in International Affairs (CEIA), in its 2016 report on conflict resolution, championed empathy as "an essential tool to resolve conflict and to ensure the sustainability of peace" (CEIA, 2016, p. 2). Together, these views converge in extolling the virtues of expressing empathy toward outgroups, adversaries, and otherwise disliked others.[12]

Evaluations of Empathizers

Given that expressing empathy toward other people—even disliked others—is encouraged, how might empathizers be evaluated? At first glance, the answer to this question seems obvious: Presumably, people who show empathy should be viewed positively, because empathy itself is highly valued. Although little work has directly examined this question, existing evidence, generally from the perspective of empathic targets, suggests that empathizers are indeed liked. Such is the case in romantic relationships (e.g., Cramer, 2003; Davis & Oathout, 1987). For example, believing that one's spouse has taken one's own perspective predicts favorable relationship outcomes (Long & Andrews, 1990). Such associations are also evident in non-romantic relationships: Patients who feel empathized with by their physicians trust their physicians more and are more likely to comply with treatment (Kim et al., 2004), and customers who feel empathized with by salespeople view these salespeople more favorably (Aggarwal et al., 2005). Because these studies were correlational and examined congenial (and often established) relationships, however, it is unclear whether the targets' positive views of empathizers are attributable to empathy *per se* or simply reflect overall relationship satisfaction.

[12] Importantly, not all scholars have a purely positive view of empathy (e.g., Bloom, 2017; Prinz, 2011; Scarry, 1996). Bloom, for instance, claims that empathy is biased and can lead to parochialism, atrocities, and immorality. Instead, he favors utilitarianism and compassion as guides for moral decision-making. Yet, the case remains that empathy is widely celebrated.

Several experiments have provided causal evidence for how empathizers are evaluated by targets in dyads of strangers. Goldstein et al. (2014), for example, examined the consequences of *perceived perspective-taking*, which they defined as the belief that another person has taken one's own perspective. Participants wrote about a personal experience (e.g., being treated unfairly by their boss), which they then shared with another ostensible participant. Participants who believed that the other participant had taken their perspective viewed that person more positively. Furthermore, this effect was mediated by participants' belief that the perspective-taker felt empathy toward them. Positive views of empathizers have also been found in relationships that are typically antagonistic: People who imagined being victims of bullying were more likely to trust and forgive the offender when they believed the offender had taken their perspective when renouncing bullying (Berndsen et al., 2018).

Notably, existing research on evaluations of empathizers has focused exclusively on how empathic targets evaluate those who have empathized with them. Given that empathic targets are likely beneficiaries of empathy, it is perhaps unsurprising that their evaluations of empathizers are positive. What remains unknown is whether empathy has evaluative implications beyond the empathic dyad. Third-party observers can form impressions of both empathizers and targets (see Figure 2.1). Indeed, empathy is often apparent in conversation speech patterns, such as the speaker's use of expressions familiar to the target and incorporation of the target's feedback (Krauss & Fussell, 1991)—information that can be readily observed by people outside the conversation. Expressions of empathy can also be directly stated. As exemplified by Bill Clinton's refrain of "I feel your pain," made famous during his 1992 U.S. presidential campaign, empathizers can express empathy toward specific people (i.e., intra-dyad targets) in a way that allows third-party observers (in Clinton's case, audience members and other potential voters) to

witness. The same is true in daily life: One might observe someone saying “I feel for you” to a friend or overhear a person say “I can put myself in your shoes” to a coworker.

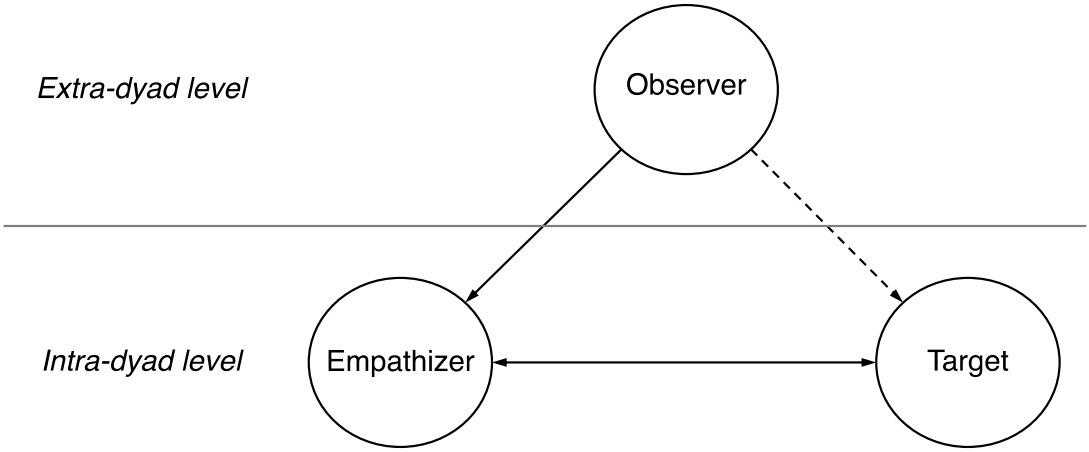

Figure 2.1. A conceptual diagram of the current research. Moving beyond the vast majority of research on empathy, which examines the effects of empathy on the empathizer or the target within a dyad (i.e., intra-dyad level), we focus on the effects of empathy beyond the dyad (i.e., extra-dyad level). Specifically, we examine how people outside an empathic dyad evaluate the empathizer (depicted as the solid arrow pointing from observer to empathizer) as a function of how observers evaluate the target (depicted as the dotted arrow pointing from observer to target).

Third-Party Observers’ Evaluations of Empathizers

How might third-party observers of expressions of empathy evaluate empathizers? On one hand, empathizers can make a positive impression on observers. For example, many credited Bill Clinton’s empathic connection with voters as a reason for his eventual win (Levine, 1993). On the other hand, as the backlash to the empathic *New York Times* profile of a white nationalist illustrates, observers’ evaluations of empathizers might not be uniformly positive and might even be negative. What remains unknown is whether, and under what conditions, expressing empathy has different consequences for third-party observers’ evaluations of empathizers.

We considered three accounts of how observers might evaluate empathizers. The first account draws from elemental approaches to impression formation and suggests that observers’ evaluations of empathizers should have the same valence as evaluations of empathy. Information

integration theories, for example, posit that evaluation of an attitude object (e.g., empathizer) is the weighted average of evaluations of relevant elements of that object (e.g., empathic expressions; Anderson, 1971): Because empathy is viewed positively, evaluations of empathizers, on average, should also be positive. Similarly, expectancy-value models of attitudes maintain that evaluation of an attitude object is a function of (a) beliefs about the attributes that characterize the object and (b) evaluations of those attributes (Fishbein & Ajzen, 1975). Thus, if one believes a person is an empathizer and evaluates empathy positively, one should evaluate the empathizer positively. This account aligns with research documenting positive evaluations of empathizers by empathic targets (Goldstein et al., 2014). Importantly, this account suggests that observers' evaluations of empathizers should *not* be calibrated to the specific target of empathy. Instead, evaluations of empathizers should simply reflect the (positive) valence of empathy itself.

The second account, which draws from balance and attribution theories (e.g., Heider, 1958; Jones & Davis, 1965), suggests that evaluations of empathizers might not be uniformly positive. Instead, observers might consider characteristics of the empathic target and form evaluations of empathizers accordingly. If observers dislike the target, for example, they should also dislike the empathizer, because the empathizer expressed affinity for the disliked target. Doing so allows observers to achieve affective balance (Heider, 1958) and to resolve conflict between the positive valence of empathy and the negative valence of the target. This proposition draws from a rich theoretical tradition on the importance of maintaining cognitive consistency (Abelson et al., 1968; Festinger, 1957; Insko, 1984; Newcomb, 1953; Osgood & Tannenbaum, 1955). According to these perspectives, inconsistencies lead to attitude change in the direction of

restoring consistency. Given a positive evaluation of empathy and a negative evaluation of the target, observers should devalue the empathizer to preserve attitudinal consistency.[13]

A third account arises from a logical integration of the first two accounts; it suggests that observers' evaluations of empathizers should be shaped by both the valence of empathy and the attitudinal consistency pressures that target characteristics impose. On this account, neither the valence of empathy nor attitudinal consistency pressures alone drive evaluations; rather, both exert forces that together shape observers' evaluations of empathizers. When the target is liked, the two forces operate in conjunction: The positive valence of empathy (i.e., "I like empathy") and the positive valence of the target (i.e., "I like the empathic target") are aligned, resulting in a positive evaluation of empathizers. When the target is disliked, however, the two forces are in opposition: The positive valence of empathy is counteracted by attitudinal consistency pressures (i.e., "That person is expressing empathy toward someone I dislike"). Because attitudinal consistency pressures should bolster evaluations of empathizers with liked targets but dampen evaluations of empathizers with disliked targets, evaluations of empathizers with disliked targets should be less positive than evaluations of empathizers with liked targets.

Both the second account and the third account posit that evaluations of empathizers should be attuned to target valence. Unlike the second account, however, the third account predicts that evaluations of empathizers with a disliked target should not fully align with the negative valence of the target, due to the positive valence of empathy acting in the opposite direction. That is, the positive effect of the valence of empathy on evaluations of empathizers

[13] These consistency-based perspectives generally suggest that people preserve consistency by adjusting evaluative elements that are easiest to change (e.g., Festinger, 1957). Thus, although updating general beliefs about empathy or changing existing evaluations of a disliked person can also allow observers to preserve consistency, both possibilities are more drastic than updating beliefs about a particular person (especially a stranger) who displays empathy and arguably less likely in many circumstances.

should be attenuated or even "canceled out" by the negative effect of the valence of the disliked target, but not fully reversed (as would be predicted by the second account). The relative strength of these two opposing forces is a key determinant of whether the positive effect of empathy is tempered or entirely eliminated when the target is disliked.

Thus, the three accounts yield different concrete predictions about how evaluations of empathizers should vary as a function of target valence. The first account posits that empathizers should be evaluated positively regardless of the target and thus predicts only a main effect of empathy and no moderation by target valence. The second account proposes that evaluations of empathizers should align with target valence and predicts a crossover interaction whereby empathizers are evaluated more positively when the target is liked but are evaluated more negatively when the target is disliked. The third account holds that evaluations of empathizers should integrate both the valence of empathy and the valence of the target; this account predicts an attenuated or even a "knockout" interaction whereby the positive effect of empathy when the target is liked is attenuated or even eliminated (but not reversed) when the target is disliked.

Overview of Experiments

Guided by these different accounts, we report seven experiments and an internal meta-analysis examining whether third-party observers' evaluations of empathizers differ based on characteristics of empathic targets. In all experiments, participants learned about an interaction in which a target disclosed a personal experience to a responder, who responded in an empathic or a non-empathic way. Participants then evaluated the responder. This paradigm reflects a common way that people observe expressions of empathy: via social interactions in verbal forms (e.g., reading online exchanges between people). More importantly, it afforded experimental control by allowing us to manipulate characteristics of both the responder and the target.

Experiment 1 examined evaluations of empathizers and the potential moderating role of target valence (i.e., whether the target is positively or negatively portrayed). Experiment 2 conceptually replicated Experiment 1 with a more realistic setup and a less extreme target valence manipulation. In Experiment 3, we changed the nature of the target's experience and explored how positive empathy (i.e., empathizing with a positive experience) affects evaluations of empathizers. Experiment 4 investigated whether the results of Experiments 1–2 could instead be explained by response positivity rather than empathy. We also assessed inferences about the responder's attitudes toward the target as a potential mediator. Our final three experiments focused on empathy toward the negatively portrayed target and examined whether a condemning (vs. empathic) response evokes more positive evaluations of the responder (Experiments 5–7), whether these effects are moderated by the gender of the characters (Experiment 6), and whether these effects may be reversed in some cases (Experiment 7).

In all experiments, participants were recruited from Amazon's Mechanical Turk (MTurk) and completed the materials online for modest remuneration. MTurk workers were eligible to participate if they lived in the U.S.; in Experiments 2–7, they were eligible only if they had not completed a previous study in this line of work. We decided *a priori* to exclude data from participants who failed any attention checks or gave identical non-neutral responses (i.e., other than 4 on 7-point scales) across all dependent variables.

We conducted power analyses to determine the target sample size for each experiment and collected data until reaching our *a priori* target sample size before analyzing data. In Experiment 1, we set a target sample size that would provide 80% power ($\alpha = .05$) to detect a small effect ($\eta_p^2 = .02$) in a 2 × 2 between-subjects design. In Experiments 2–7, we set conservative target sample sizes based on power analyses that used effect size estimates

observed in our previous experiments. We report sample sizes and data exclusions in the main text; participant details and power for each experiment appear in the Supplemental Materials.

For each experiment, we report all conditions, manipulations, and key dependent measures of interest. All manipulations in all experiments were successful; details appear in the Supplemental Materials. We distinguish between planned and unplanned (exploratory) data analyses, and we note departures from planned data analyses where appropriate.

Experiment 1

Experiment 1 was our first test of whether evaluations of empathizers depend on the valence of the target. Participants read about an interaction between Ann (target) and Beth (responder), who were meeting for the first time. They learned that Ann, who worked for either a children's hospital (positive target) or a white supremacist group (negative target), disclosed a stressful experience, and that Beth responded in an empathic or non-empathic way. Our three accounts yield different predictions. The first account predicts only a main effect of response type and no moderation by target valence. The second account predicts a crossover interaction whereby the empathic response results in more positive evaluations of the responder when target valence is positive but more negative evaluations when target valence is negative. The third account predicts an attenuated interaction whereby the positive effect of the empathic response in the positive target condition is weaker (and possibly eliminated) in the negative target condition.

Method

Participants. Participants were 464 MTurk workers. Based on our *a priori* exclusion criteria, we excluded $n = 89$ for failing the attention check on Beth's response, $n = 52$ for failing

the attention check on Ann's employer, and $n = 7$ for giving identical non-neutral responses to the dependent variables. The final sample was $N = 336$.[14]

Materials and procedure. Participants were randomly assigned to one of the 2 (response type: empathic vs. non-empathic) × 2 (target valence: positive vs. negative) between-subjects conditions. As part of a study on "first impressions," participants learned about an interaction between two people, Beth and Ann. Participants saw an ostensible business card belonging to Ann; it included her name, contact information, and, critically, her employer. In the *positive target* condition, Ann did event planning and outreach for St. Jude Children's Research Hospital. In the *negative target* condition, Ann did event planning and outreach for Aryan Nations (a white supremacist group; see Figure 2.2). To ensure that participants understood the mission of Ann's employer, organization slogans appeared on the business cards ("Finding Cures / Saving Children" for St. Jude Children's Research Hospital, "White People Awake / Save Our Great Race" for Aryan Nations).[15] Participants then reported their first impression of Ann (1 = *very negative*, 4 = *neutral*, 7 = *very positive*) as a manipulation check on target valence.

Figure 2.2. Stimuli used in Experiment 1 to manipulate target valence. In the positive target condition (left), the target (Ann Russell) works for St. Jude Children's Research Hospital; in the

[14] Due to a programming oversight, we did not collect information on participant gender and age in Experiment 1. We report participant gender and age for all other experiments.

[15] Both slogans are real. "Finding cures. Saving children." is indeed the slogan of St. Jude Children's Research Hospital. "White people awake, save our great race" is commonly associated with the Hammerskin Nation, another white supremacist group (Tenold, 2018). We decided to use Aryan Nations because it is more well-known.

negative target condition (right), the target works for Aryan Nations. Fictitious contact information (redacted here) appeared on the business card.

Next, participants read an excerpt of an interaction between Beth and Ann, who were meeting for the first time. Beth had just learned about Ann's job, and Ann was telling Beth about a recent stressful experience. All participants read the following statement from Ann:

> "I'm feeling really stressed. I'm organizing an event, and my team is expecting a large attendance. I've been having trouble with the logistics of it, and the date of the event was recently delayed because we did not hear back from the city council in time. The stress has affected my sleep, and I've been feeling awful because of it."

Participants then saw Beth's response. In the *empathic response* condition, Beth said, "I feel for you—I can really put myself in your shoes in this situation. When is the event taking place?" In the *non-empathic response* condition, Beth said, "Okay, I see. When is the event taking place?"

Following the excerpt, as an attention check, participants identified Beth's response to Ann from a list ("I can really put myself in your shoes in this situation," "Okay, I see," "I do not understand your situation," and *none of the above*). They then completed the primary dependent measures assessing evaluations of Beth by indicating how much they *liked*, *respected*, *trusted*, and *would like to be friends with* Beth (1 = *not at all*, 7 = *very much*), and how *understanding*, *kind*, *cold* (reverse-coded), and *caring* Beth was (1 = *not at all* _____, 7 = *very* _____, with _____ as the trait word). Participants then completed an exploratory measure,[16] a manipulation check on response empathy ("To what extent do you think Beth empathized with Ann?" 1 = *not at all*, 7 = *very much*), and an attention check on Ann's employer ("St. Jude Children's Research Hospital," "Aryan Nations," "Pacific Gas and Electric Company," and *no work information of*

[16] In this and all subsequent experiments, we included an exploratory item assessing beliefs about the similarity between Beth and Ann. Because this variable was not central to our research questions, we report it here for transparency but do not discuss it further. Exploratory analyses on this item appear in the Supplemental Materials.

Ann was given). Lastly, they answered an open-ended question on their reaction to the interaction and completed demographic questions.

Results

Data reduction. To reduce the dimensions of our primary dependent variables, we conducted an exploratory factor analysis (EFA) using promax rotation in R (R Core Team, 2019) and arrived at a two-factor solution, $\chi^2(13) = 22.88$, $p = .043$.[17] Four items loaded onto the first factor, which we interpreted as *respect/liking*; the other four items loaded onto the second factor, which we interpreted as *warmth* (see Figure 2.3). Each item loaded onto its primary factor at higher than $\lambda = .70$ and the other factor at lower than $\lambda = .25$. Solutions with three or more factors did not have theoretically sensible structures or item loadings on any additional factors above $\lambda = .35$, and the solution with one factor did not describe the data well, $\chi^2(20) = 283.26$, $p < .001$; thus, we retained our two-factor solution, which accounted for 66% of the total variance (Factor 1 = 35.6%; Factor 2 = 30.0%). Based on this factor structure and the comparable item loadings within each factor, we calculated the mean ratings of the first four items as a respect/liking composite ($\alpha = .95$) and the mean ratings of the last four items as a warmth composite ($\alpha = .90$). Although the respect/liking and warmth composites were highly correlated ($r = .79$, $p < .001$), EFA suggested that they were best considered as distinct dimensions, so we conducted our primary analyses on these composites separately.

[17] Following Flora and Flake's (2017) recommendations, we verified that our interpretation of the factors was consistent across several oblique rotations and estimation methods.

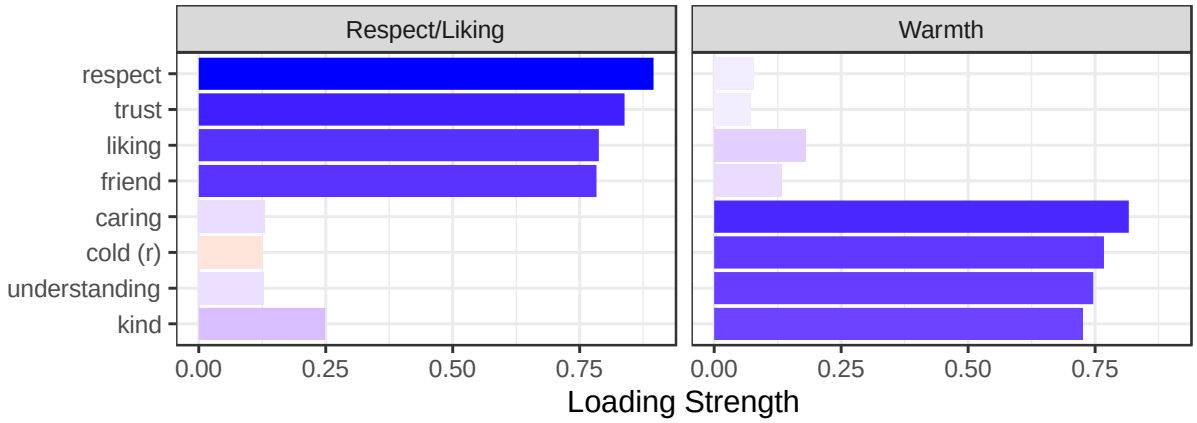

Figure 2.3. Results from the EFA on the primary dependent variables in Experiment 1. The *x* axis depicts the absolute loading strength of an item on the factor indicated in the panel headings. Blue horizontal bars are positive factor loadings; red horizontal bars are negative factor loadings.

Respect/liking. A 2 (response type: empathic vs. non-empathic) × 2 (target valence: positive vs. negative) between-subjects analysis of variance (ANOVA) on respect/liking revealed that participants respected/liked Beth more when she gave an empathic (vs. non-empathic) response, $F(1, 332) = 14.37$, $p < .001$, $\eta_p^2 = .04$, $CI_{90\%}$ [.01, .08], and when Ann was positively (vs. negatively) portrayed, $F(1, 332) = 48.62$, $p < .001$, $\eta_p^2 = .13$, $CI_{90\%}$ [.08, .18]. More importantly, the response type × target valence interaction was significant, $F(1, 332) = 5.40$, $p = .021$, $\eta_p^2 = .02$, $CI_{90\%}$ [.001, .05]. When Ann was positively portrayed, participants respected/liked Beth more when she gave an empathic (vs. non-empathic) response ($M = 5.33$, $SD = 1.01$ vs. $M = 4.46$, $SD = 1.15$), $F(1, 332) = 20.12$, $p < .001$, $\eta_p^2 = .06$, $CI_{90\%}$ [.02, .10]. When Ann was negatively portrayed, however, respect/liking for Beth did not significantly differ by response type ($M = 4.01$, $SD = 1.73$ vs. $M = 3.80$, $SD = 1.12$), $F(1, 332) = 1.01$, $p = .317$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .02] (Figure 2.4, left panel).

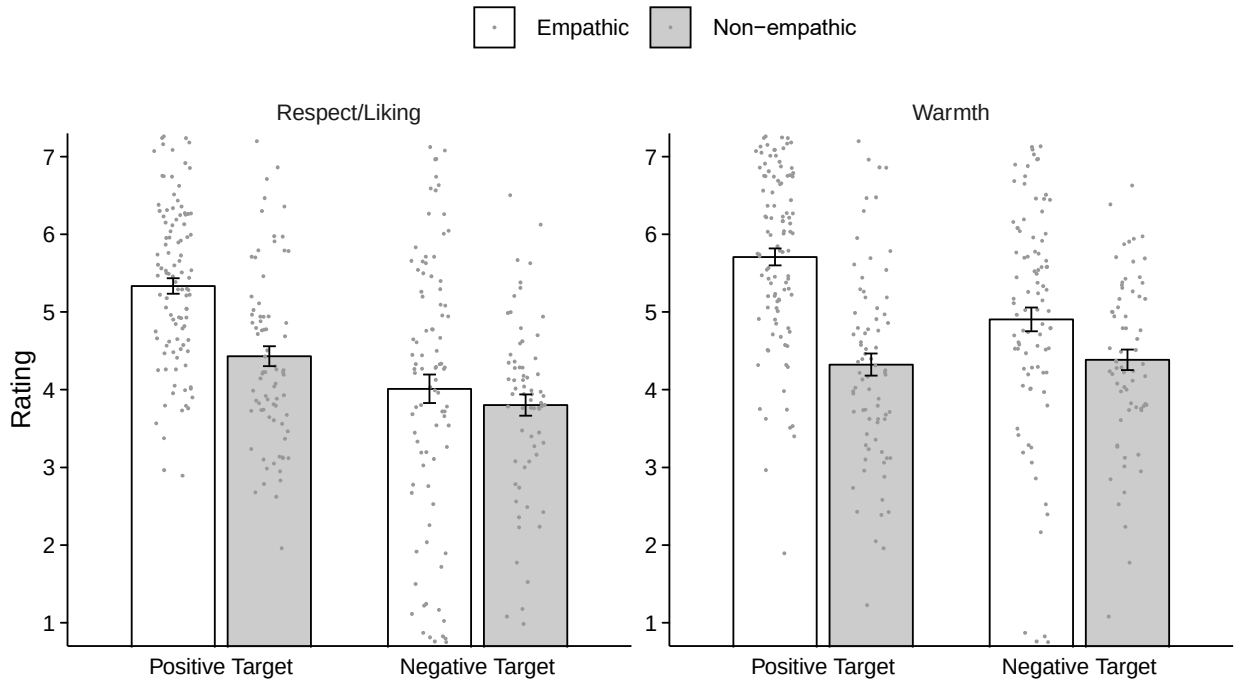

Figure 2.4. Ratings of Beth on respect/liking and warmth by response type and target valence in Experiment 1. Error bars depict ±1 standard errors; dots depict jittered individual data points.

Warmth. An identical 2 × 2 ANOVA on warmth revealed that participants rated Beth as warmer when she gave an empathic (vs. non-empathic) response, $F(1, 332) = 46.80$, $p < .001$, $\eta_p^2 = .12$, $CI_{90\%}$ [.07, .18], and when Ann was positively (vs. negatively) portrayed, $F(1, 332) = 8.07$, $p = .005$, $\eta_p^2 = .02$, $CI_{90\%}$ [.004, .06]. More importantly, the response type × target valence interaction was significant, $F(1, 332) = 9.23$, $p = .003$, $\eta_p^2 = .03$, $CI_{90\%}$ [.01, .06]. When Ann was positively portrayed, participants rated Beth as warmer when she gave an empathic (vs. non-empathic) response ($M = 5.71$, $SD = 1.10$ vs. $M = 4.36$, $SD = 1.27$), $F(1, 332) = 52.54$, $p < .001$, $\eta_p^2 = .14$, $CI_{90\%}$ [.08, .19]. Unlike the results for respect/liking, even when Ann was negatively portrayed, participants still rated Beth as warmer when she gave an empathic (vs. non-empathic) response ($M = 4.90$, $SD = 1.45$ vs. $M = 4.38$, $SD = 1.08$), though this effect was significantly smaller, $F(1, 332) = 6.75$, $p = .010$, $\eta_p^2 = .02$, $CI_{90\%}$ [.003, .05] (Figure 2.4, right panel).

Discussion

Experiment 1 provides initial evidence that evaluations of empathizers depend on to whom empathy is shown. Participants respected/liked the empathizer more when the target was positively portrayed, but not when the target was negatively portrayed. Participants rated the empathizer as warmer overall, but this effect was stronger when the target was positively (vs. negatively) portrayed. Experiment 1 also provides initial evidence that respect/liking and warmth reflect two related but distinct dimensions along which participants evaluated the responder.

The study materials contained several ambiguities, however, that might have contributed to these results. Although the instructions explicitly stated that Beth and Ann were meeting for the first time, some participants might have assumed that they knew each other beforehand. If so, perhaps the observed effects are due, in part, to participants' beliefs about the relationship between Beth and Ann (e.g., Beth associates with a white supremacist, so Beth is not a good person), rather than Beth's response to Ann. Furthermore, because the target valence manipulation appeared before the dialogue, participants might have assumed that Beth did not know that Ann worked for Aryan Nations or what Aryan Nations is.

To address these ambiguities, we conducted a conceptual replication of Experiment 1 (see Experiment S1 in the Supplemental Materials). We extended the dialogue between Beth and Ann to clarify that (a) they did not know each other beforehand, and that (b) Beth learned, via Ann's self-disclosure to her, what organization Ann worked for and understood its mission. Results largely replicated those of Experiment 1.

Together, Experiments 1 and S1 indicate that evaluations of empathizers depended on target valence. When the target was positively portrayed, empathizers were respected/liked more and were rated as warmer than non-empathizers; when the target was negatively portrayed, empathizers were still rated as warmer, but they were no longer respected/liked more. In

Experiment 2, we used a different paradigm and target valence manipulation to test the generalizability of these findings. We also modified the responses to rule out a potential confound: Beth's question "When is the event taking place?" might have implied interest in attending the event; thus, we removed this question from all conditions in Experiment 2.

Experiment 2

Experiment 1 used a vignette-based paradigm in which the target was portrayed as working for either a children's hospital or a white supremacist organization. Although the strength of this manipulation[18] helped maximize the statistical power of our experimental design (Ledgerwood, 2019), it is possible that the findings in Experiment 1 depend on this particular manipulation and would not replicate with a less extreme target valence manipulation. Furthermore, although the vignettes resemble some real-world scenarios (e.g., reading about an empathic exchange between two people on social media), participants might have treated the interaction as a hypothetical scenario and might have reacted differently if they believed the interaction was real. Therefore, in Experiment 2, we tested the generalizability of the key findings from Experiment 1 by using a less extreme target valence manipulation (target holding pro-vaccination vs. anti-vaccination beliefs) and presenting the interaction as part of an ostensible, in-person study. These changes allowed us to test if the findings from Experiment 1 are limited to the particular manipulation and paradigm or are broader in scope.

Method

Participants. We publicly pre-registered our analysis plan on AsPredicted (https://aspredicted.org/blind.php?x=fr6rn9). Participants were 614 MTurk workers (49%

[18] The effect size of the manipulation check was $d = 2.97$, $CI_{95\%}$ [2.66, 3.29] (see Supplemental Materials).

female, 51% male; M_{age} = 37.9, SD_{age} = 12.5). We excluded n = 88 from data analyses based on our *a priori* exclusion criteria. The final sample was N = 526.

Materials and procedure. Participants were randomly assigned to one of the 2 (response type: empathic vs. non-empathic) × 2 (target valence: positive vs. negative) between-subjects conditions. Similar to Experiment 1, participants learned about an interaction between Beth and Ann, who were meeting for the first time. Unlike Experiment 1, participants learned that the interaction between Beth and Ann was ostensibly recorded and transcribed as part of an in-person study previously conducted in the lab, and that their real names had been replaced with pseudonyms for purposes of anonymity. Participants then learned that Beth and Ann had filled out a survey prior to their interaction and shared their answers with each other. All participants were assigned to read Ann's ostensible answer to the survey question, "What is an issue you care about?" In the *positive target* condition, Ann's answer implied that she held pro-vaccination beliefs; in the *negative target* condition, Ann's answer implied that she held anti-vaccination beliefs (see Figure 2.5 for exact wording). Participants then completed the same manipulation check from Experiment 1 by reporting their impression of Ann.[19]

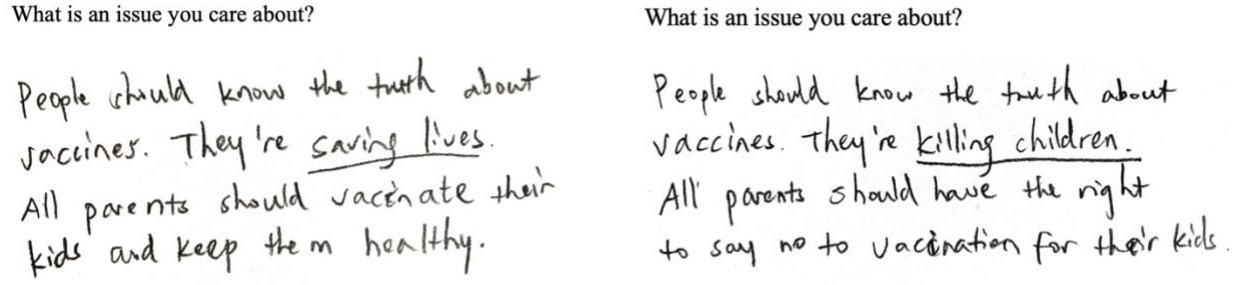

Figure 2.5. Stimuli used in Experiment 2 to manipulate target valence. In the positive target condition (left), the target (Ann) expressed pro-vaccination beliefs; in the negative target

[19] This cover story appeared convincing to participants, most of whom commented on the interaction in their open-ended responses at the end of the experiment (e.g., what they would have said to Ann, wanting to know what happened after the interaction). Three participants expressed suspicion about the veracity of our cover story, but excluding their responses did not change the significance of any result. Following our pre-analysis plan, we retained their data in the analyses reported below.

condition (right), the target expressed anti-vaccination beliefs. To enhance the perceived authenticity of the stimuli, both answers were handwritten and contained an ambiguous typo in "vaccinate"/ "vaccination."

Next, participants read an excerpt of the ostensible interaction between Beth and Ann. Text for Ann's statement in the positive target conditions appears below; in the negative target conditions, the organization name was "Stop Mandatory Vaccination":

> "So yeah, I work for an organization called Vaccinate Your Family, and I'm putting together an event for them. My team is expecting a large attendance, but I've been having a lot of trouble with the logistics of it, and the date of the event was recently delayed because we did not hear back from the city council in time. I've been under a lot of stress, and it is really overwhelming. I'm not sleeping well, and I've been feeling awful because of it."

In the *empathic response* condition, Beth responded, "I feel for you—I can really put myself in your shoes in this situation." In the *non-empathic response* condition, Beth responded, "Okay, I see."

Participants then completed the same dependent measures from Experiment 1, a manipulation check on Ann's affect ("How positive did Ann feel when she told Beth about her recent experience at work?" 1 = *very negative,* 7 = *very positive*), an exploratory measure on the positivity of Beth's response ("How positive was Beth's response to Ann's disclosure about her experience at work?" 1 = *very negative,* 7 = *very positive*), and an exploratory measure on general attitudes toward vaccines ("In general, what are your views on vaccinations?" 1 = *very negative,* 4 = *neutral/mixed feelings,* 7 = *very positive*).

Results

The target valence manipulation was successful: Participants evaluated Ann more positively when she was portrayed as pro- versus anti-vaccination (M = 5.66, SD = 1.19 vs. M = 2.80, SD = 1.74), $t(442) = 21.94$, $p < .001$, $d = 1.94$, $CI_{95\%}$ [1.73, 2.15]. As expected, this target

valence manipulation was considerably weaker than the manipulation used in Experiment 1 (see Footnote 7). Results of other manipulation checks are available in the Supplemental Materials.

Factor analysis. To confirm the factor structure from Experiment 1, we conducted a confirmatory factor analysis (CFA) in R using the *lavaan* package (Rosseel, 2012). Drawing from the EFA solution in Experiment 1, we specified a model with two latent factors; four items (*like*, *respect*, *trust*, and *friends*) loaded onto the first factor (*respect/liking*), and the other four items (*understanding*, *kind*, *cold* [reverse-coded], and *caring*) loaded onto the second factor (*warmth*). Because factor loadings of all items on their non-primary factors were low in the EFA solution in Experiment 1, we specified no cross-loadings in the CFA. This two-factor model fit the data well, $\chi^2(19) = 102.63$, $p < .001$, *RMSEA* = 0.09, *CFI* = 0.98, *TLI* = 0.97, with all factor loadings higher than $\lambda = .60$. The two-factor model also fit the data better than a one-factor model in which all items loaded onto a single factor, $\Delta\chi^2(1) = 430.82$, $p < .001$. Thus, we confirmed the factor structure from Experiment 1. As in Experiment 1, we calculated the mean ratings of items for respect/liking ($\alpha = .94$) and warmth ($\alpha = .90$) as composites and conducted the primary analyses on these composites.[20]

Respect/liking. A 2 (response type) × 2 (target valence) between-subjects ANOVA on respect/liking revealed that participants respected/liked Beth more when she gave an empathic (vs. non-empathic) response, $F(1, 522) = 66.10$, $p < .001$, $\eta_p^2 = .11$, $CI_{90\%}$ [.07, .16]. The target valence main effect was not significant, $F(1, 522) = 0.84$, $p = .359$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .01]. More importantly, the response type × target valence interaction was significant, $F(1, 522) =$

[20] Following Flake et al.'s (2017) recommendations for ongoing construct validation, we confirmed the factor structure observed here with CFA in subsequent experiments, all of which supported the same two-factor structure (i.e., it fit the data well and provided substantially better fit than a one-factor model, which fit the data poorly). We only report the internal consistencies of the composite scores in Experiments 3–7; the full set of CFA results are available in the Supplemental Materials.

33.03, $p < .001$, $\eta_p^2 = .06$, $CI_{90\%}$ [.03, .09]. When Ann was positively portrayed, participants respected/liked Beth more when she gave an empathic (vs. non-empathic) response ($M = 5.11$, $SD = 1.04$ vs. $M = 3.69$, $SD = 1.30$), $F(1, 522) = 99.69$, $p < .001$, $\eta_p^2 = .16$, $CI_{90\%}$ [.12, .21]. When Ann was negatively portrayed, participants respected/liked Beth more when she gave an empathic (vs. non-empathic) response ($M = 4.62$, $SD = 1.29$ vs. $M = 4.38$, $SD = 1.04$), but this effect was smaller and not significant, $F(1, 522) = 2.75$, $p = .098$, $\eta_p^2 = .01$, $CI_{90\%}$ [.00, .02] (Figure 2.6, left panel).

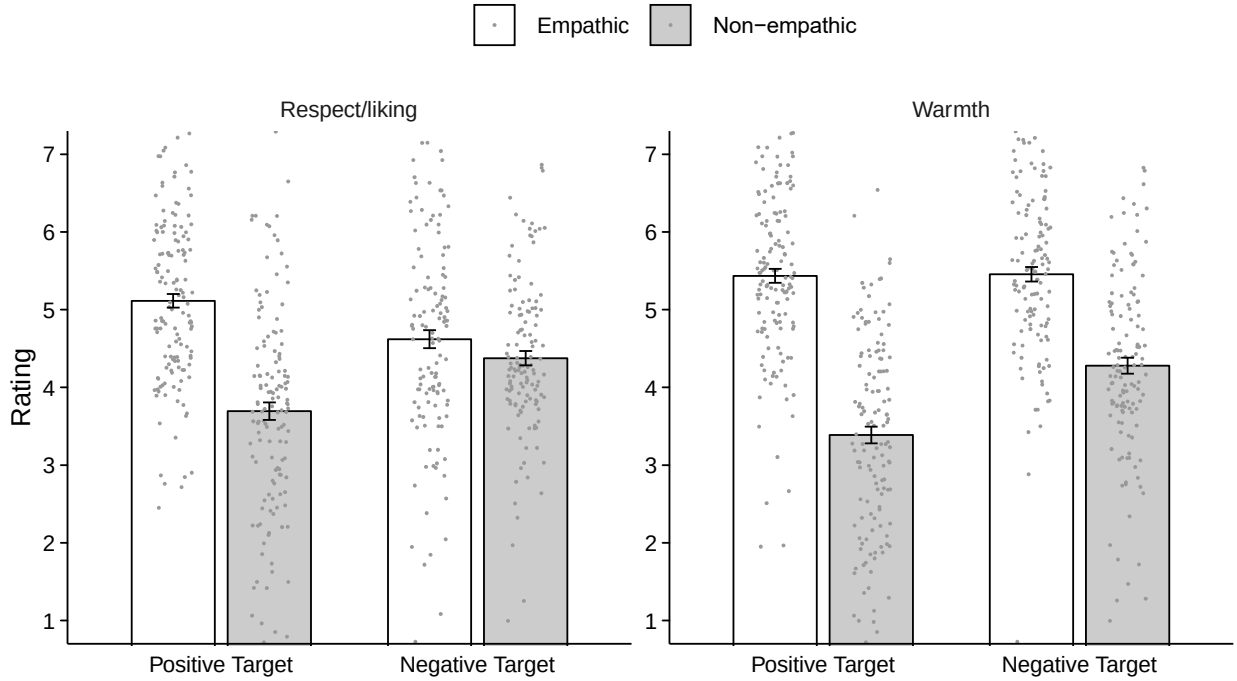

Figure 2.6. Ratings of Beth on respect/liking and warmth by response type and target valence in Experiment 2. Error bars depict ±1 standard errors; dots depict jittered individual data points.

Warmth. An identical 2 × 2 ANOVA on warmth revealed that participants rated Beth as warmer when she gave an empathic (vs. non-empathic) response, $F(1, 522) = 264.72$, $p < .001$, $\eta_p^2 = .34$, $CI_{90\%}$ [.28, .38], and when Ann was *negatively* (vs. positively) portrayed, $F(1, 522) = 21.20$, $p < .001$, $\eta_p^2 = .04$, $CI_{90\%}$ [.02, .07]. More importantly, the response type × target valence interaction was significant, $F(1, 522) = 19.31$, $p < .001$, $\eta_p^2 = .04$, $CI_{90\%}$ [.01, .07]. When Ann

was positively portrayed, participants rated Beth as warmer when she gave an empathic (vs. non-empathic) response (M = 5.44, SD = 1.06 vs. M = 3.39, SD = 1.24), $F(1, 522) = 221.06$, $p < .001$, $\eta_p^2 = .30$, $CI_{90\%}$ [.25, .35]. When Ann was negatively portrayed, participants still rated Beth as warmer when she gave an empathic (vs. non-empathic) response (M = 5.46, SD = 1.05 vs. M = 4.28, SD = 1.18), but this effect was smaller, $F(1, 522) = 68.19$, $p < .001$, $\eta_p^2 = .12$, $CI_{90\%}$ [.08, .16] (Figure 2.6, right panel).

Discussion

Experiment 2 replicated the key findings from Experiment 1. Using a less extreme target valence manipulation and a more realistic setup, we again found that evaluations of empathizers depended on target valence. The interaction pattern was largely the same as that in Experiment 1: Participants respected/liked the responder more when she responded empathically to a positively portrayed target, but not when she responded to a negatively portrayed target. Participants also rated the responder as warmer when she responded empathically, but this effect was smaller when the target was negatively portrayed. The sizes of the interaction effects were comparable with those in Experiment 1, perhaps because the enhanced realism of the paradigm compensated for the weaker target valence manipulation.

Notably, we observed no evidence of backfiring in either experiment: Participants did not respect/like empathizers of a negatively portrayed target *less*. If anything, the pattern of results for the negatively portrayed target was in the same direction descriptively, with participants respecting/liking empathizers slightly more. This pattern is consistent with our "opposing forces" account, which suggests that the non-reversal in respect/liking when the target was negatively portrayed resulted from the valence of empathy and the attitudinal consistency pressures operating in opposition. On one hand, empathy is generally liked; it is also the default response

to a person experiencing negative affect (McAuliffe et al., 2020). Thus, it is likely that an empathic response was both expected and viewed positively when the target experienced stress. On the other hand, this positive view might be attenuated by attitudinal consistency pressures toward viewing Beth negatively because she expressed empathy for a white supremacist. If the null effect of response type on respect/liking in the negative target conditions was due to the two forces—the positive view of expressing empathy in response to negative affect (which should increase respect/liking) and the attitudinal consistency pressures (which should decrease respect/liking)—canceling each other out, then shifting the relative strength of those forces should change the results.

We explored this possibility in Experiment 3. We reasoned that when a negative target discloses a *positive* experience, the influence of the valence of empathy on evaluations of the empathizer should diminish, because empathy may no longer be the default, expected response. In this way, responding empathically to a positive experience should be especially diagnostic of the responder's values as a person (i.e., as someone who responds in an active–constructive manner to a white supremacist's positive disclosure; see Gable & Reis, 2010), thereby enhancing attitudinal consistency pressures. Consequently, empathizers (vs. non-empathizers) might be evaluated more negatively when a disliked target discloses a positive experience.

Experiment 3

Experiment 3 tested whether target valence moderates evaluations of empathizers (vs. non-empathizers) when the target discloses a positive rather than a stressful experience. Positive empathy refers to understanding and sharing others' positive emotions (Morelli et al., 2015a). It is closely related to negative empathy (i.e., sharing and understanding others' negative emotions; Gable et al., 2006). Yet, positive and negative empathy differ in the valence of the shared

experience (Morelli et al., 2015b). More central to the goal of Experiment 3, examining positive empathy allowed us to test our "opposing forces" account, which accommodates the results of Experiments 1–2 but predicts a different interaction pattern here.

Specifically, we expected a crossover response type × target valence interaction whereby the effect of response type on respect/liking for Beth when Ann was positively portrayed would reverse when Ann was negatively portrayed. Unlike Experiments 1–2, a normative expectation of empathy was less likely to be operating here, given that Ann disclosed a positive experience. Thus, we predicted that participants would respect/like Beth *less* when she gave an empathic response to negatively portrayed Ann. We did not have *a priori* predictions for warmth. Because the effect sizes of the interactions were comparable in Experiments 1–2, we returned to the paradigm from Experiment 1.

Method

Participants. Participants were 507 MTurk workers (52% female, 44% male, 4% no gender information; M_{age} = 37.4, SD_{age} = 12.6). We excluded n = 91 from data analyses based on our *a priori* exclusion criteria. The final sample was N = 416.

Materials and procedure. This experiment was identical to Experiment 1, the only difference being that participants read that Ann told Beth about a recent positive experience (instead of a stressful experience):

> "Things have been going really well lately. I recently organized an event, and it was a huge success. A lot of people showed up to participate, and we received a large anonymous donation, which is going to make my job so much easier in the future. On top of that, I just found out that I got a raise!"

Next, participants saw Beth's response. In the *empathic response* condition, Beth said, "Good for you! I can imagine how excited you must feel" (see Reis et al., 2010, for a similar positive empathy expression). In the *non-empathic response* condition, Beth said, "Okay, I see."

Participants completed the same set of measures from Experiments 1–2 and the same manipulation check on Ann's affect from Experiment 2.

Results

Respect/liking. A 2 (response type) × 2 (target valence) between-subjects ANOVA on respect/liking ($\alpha = .95$) revealed that participants respected/liked Beth more when she gave an empathic (vs. non-empathic) response, $F(1, 412) = 28.20$, $p < .001$, $\eta_p^2 = .06$, $CI_{90\%}$ [.03, .11], and when Ann was positively (vs. negatively) portrayed, $F(1, 412) = 12.81$, $p < .001$, $\eta_p^2 = .03$, $CI_{90\%}$ [.01, .06]. More importantly, the expected crossover response type × target valence interaction was significant, $F(1, 412) = 91.27$, $p < .001$, $\eta_p^2 = .18$, $CI_{90\%}$ [.13, .24]. When Ann was positively portrayed, participants respected/liked Beth more when she gave an empathic (vs. non-empathic) response ($M = 5.24$, $SD = 0.92$ vs. $M = 3.36$, $SD = 1.38$), $F(1, 412) = 122.17$, $p < .001$, $\eta_p^2 = .23$, $CI_{90\%}$ [.17, .28]. When Ann was negatively portrayed, however, participants respected/liked Beth *less* when she gave an empathic (vs. non-empathic) response ($M = 3.58$, $SD = 1.64$ vs. $M = 4.12$, $SD = 1.14$), $F(1, 412) = 8.22$, $p = .004$, $\eta_p^2 = .02$, $CI_{90\%}$ [.004, .05] (Figure 2.7, left panel).

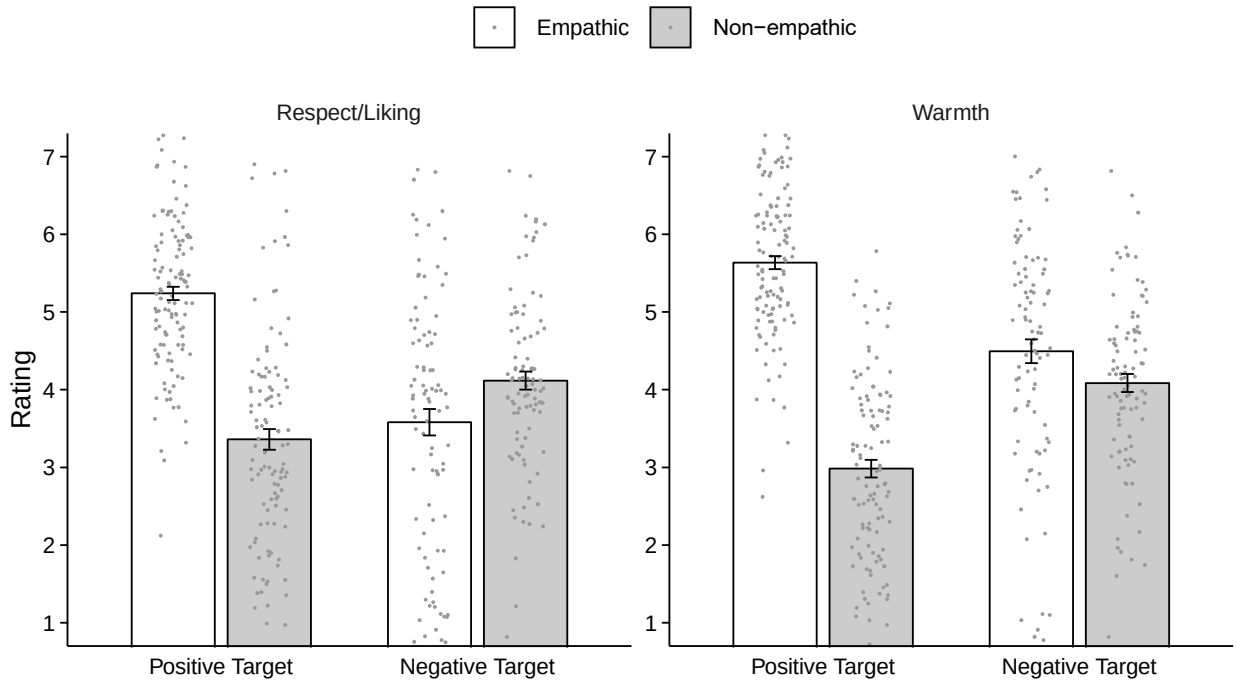

Figure 2.7. Ratings of Beth on respect/liking and warmth by response type and target valence in Experiment 3. Error bars depict ±1 standard errors; dots depict jittered individual data points.

Warmth. An identical 2 × 2 ANOVA on warmth (α = .91) revealed that participants rated Beth as warmer when she gave an empathic (vs. non-empathic) response, $F(1, 412) = 173.30$, $p < .001$, $\eta_p^2 = .30$, $CI_{90\%}$ [.24, .35]. Unlike the results for respect/liking, however, the target valence main effect was not significant, $F(1, 412) = 0.04$, $p = .835$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .00]. The response type × target valence interaction was significant, $F(1, 412) = 92.89$, $p < .001$, $\eta_p^2 = .18$, $CI_{90\%}$ [.13, .24]. When Ann was positively portrayed, participants rated Beth as warmer when she gave an empathic (vs. non-empathic) response ($M = 5.64$, $SD = 0.91$ vs. $M = 2.99$, $SD = 1.19$), $F(1, 412) = 287.51$, $p < .001$, $\eta_p^2 = .41$, $CI_{90\%}$ [.35, .46]. When Ann was negatively portrayed, participants still rated Beth as warmer when she gave an empathic (vs. non-empathic) response ($M = 4.49$, $SD = 1.48$ vs. $M = 4.09$, $SD = 1.13$), though this effect was smaller, $F(1, 411) = 5.67$, $p = .018$, $\eta_p^2 = .01$, $CI_{90\%}$ [.001, .04] (Figure 2.7, right panel).

Discussion

Experiment 3 tested whether the results observed in Experiments 1–2 would change when the target disclosed a positive experience. As before, target valence moderated both respect/liking for and warmth toward the responder (Beth), but the pattern of moderation for respect/liking was different here. Participants respected/liked Beth more when she gave an empathic response to positively portrayed Ann, but they respected/liked Beth *less* when she gave an empathic response to negatively portrayed Ann. In contrast, participants still rated Beth as warmer when she gave an empathic response to negatively portrayed Ann (though this effect was smaller than that in the positive target condition).

These results suggest that target valence moderates evaluations of empathizers, regardless of whether the target's experience is negative or positive. The pattern of moderation for respect/liking is consistent with our "opposing forces" account: Because experiencing a positive event (e.g., an accomplishment) dampens the expectation of an empathic response, this dampened expectation, in turn, should both diminish the influence of the valence of empathy and exert greater attitudinal consistency pressures on evaluations of the empathizer when the target is negatively portrayed. A different interaction pattern emerged for warmth; we revisit this observation in the General Discussion.

One limitation of this experiment is that our response type manipulation might have inadvertently manipulated more than empathy. Specifically, participants might have interpreted the first part of the empathic response, "good for you," as indicative of Beth's approval of Ann's work. Although participants rated the empathic response as comparably empathic across target valence (M = 5.55, SD = 1.07 vs. M = 5.63, SD = 1.36), $t(170) = 0.46$, $p = .647$, $d = 0.07$, $CI_{95\%}$ [-0.21, 0.34], it is possible that the simple main effect of response type on respect/liking in the negative target condition partially reflects what participants inferred about Beth based on her

positive response to someone who works for a children's hospital versus a white supremacist organization. In Experiment 4, we tested the role of positivity in driving the effect of response type on evaluations of empathizers.

Experiment 4

Thus far, we have examined evaluations of empathic versus non-empathic responders. It is possible, however, that it is not empathy *per se* that is driving these effects, but rather response *positivity*. Our exploratory measure on response positivity in Experiments 2 and 3 suggested that the empathic response is undeniably more positive than the non-empathic response, *d*s = 1.32–2.60. Empathic responses naturally tend to be positive (indeed, it is difficult to imagine an ecologically valid response that is both empathic and neutral). Yet, if a difference in positivity between the empathic and the non-empathic responses underlies the effects, they should disappear when the responses are equated on positivity. In Experiment 4, we manipulated response positivity to test whether the results observed in Experiments 1–3 would still emerge (moderation-of-process design; Spencer et al., 2005).

We also examined how these results might be related to participants' inferences about the *responder's* attitudes toward the target. We reasoned that if participants infer that the responder's evaluation of the target differs from their own evaluation of the target, then they should evaluate the responder less positively. We assessed inferences about Beth's attitudes toward Ann as a potential mediator of the response type × target valence interaction on both respect/liking and warmth (measurement-of-mediation design; Spencer et al., 2005).

Method

Participants. We publicly pre-registered our analysis plan on AsPredicted (http://aspredicted.org/blind.php?x=xu9ur5). Participants were 838 MTurk workers (58% female,

42% male; M_{age} = 39.6, *SD* = 12.4). We excluded *n* = 98 from data analyses based on our *a priori* exclusion criteria. The final sample was *N* = 740.

Materials and procedure. Participants were randomly assigned to one of the 3 (response type: positive empathic vs. positive non-empathic vs. neutral non-empathic) × 2 (target valence: positive vs. negative) between-subjects conditions. The target valence manipulation was identical to previous experiments. After the target valence manipulation and its manipulation check, participants saw the same instructions and Ann's statement from Experiment 1.

Participants then saw Beth's response. In the *empathic response* condition, Beth said, "I feel for you—I can really put myself in your shoes in this situation." In the *positive non-empathic response* condition, Beth said, "Just stay positive! Life is better when we look on the bright side." In the *neutral non-empathic response* condition, Beth said, "Okay, I see."[21]

Next, participants completed the same dependent measures and attention checks (with updated response options to reflect the current response type manipulation) as before. We used the same manipulation checks from Experiments 2–3. We measured inferences about Beth's attitudes toward Ann with three items: Participants indicated how much they agreed that Beth liked Ann, felt positive toward Ann, and had an unfavorable opinion of Ann (reverse-coded; 1 = *strongly disagree*, 7 = *strongly agree*; α = .89).

Results

We followed our pre-analysis plan for all planned analyses reported below. We also conducted several exploratory analyses, which we report as such below.

[21] We conducted a pilot study (*N* = 201) in which participants evaluated Beth's different responses without information about Ann. Results confirmed that the empathic and the positive non-empathic responses were comparably positive, and that both responses were more positive than the neutral non-empathic response. The empathic response was also more empathic than both the positive non-empathic response and the neutral non-empathic response. We report details of the pilot study in the Supplemental Materials.

Respect/liking. A 3 (response type: empathic vs. positive non-empathic vs. neutral non-empathic) × 2 (target valence) between-subjects ANOVA on respect/liking (α = .96) revealed a response type main effect, $F(2, 734) = 23.68$, $p < .001$, $\eta_p^2 = .06$, $CI_{90\%}$ [.03, .09]. The target valence main effect was also significant: Participants respected/liked Beth more when Ann was positively (vs. negatively) portrayed, $F(1, 734) = 5.24$, $p = .022$, $\eta_p^2 = .01$, $CI_{90\%}$ [.001, .02]. More importantly, the response type × target valence interaction was significant, $F(2, 734) = 17.79$, $p < .001$, $\eta_p^2 = .05$, $CI_{90\%}$ [.02, .07]. Planned contrasts in the positive target condition indicated that participants respected/liked Beth more when she gave an empathic (vs. positive non-empathic) response ($M = 5.06$, $SD = 1.11$ vs. $M = 4.18$, $SD = 1.53$), $t(734) = 5.52$, $p < .001$, $d = 0.67$, $CI_{95\%}$ [0.42, 0.93], and when she gave an empathic (vs. neutral non-empathic) response ($M = 3.60$, $SD = 1.20$), $t(734) = 8.76$, $p < .001$, $d = 1.12$, $CI_{95\%}$ [0.83, 1.40]. Participants also respected/liked Beth more when she gave a positive non-empathic (vs. neutral non-empathic) response, $t(734) = 3.40$, $p < .001$, $d = 0.44$, $CI_{95\%}$ [0.18, 0.70] (Figure 2.8, left panel).

Though not planned, we also explored whether respect/liking for Beth differed among the three response types in the negative target condition. Post-hoc pairwise comparisons indicated that participants respected/liked Beth more when she gave a positive non-empathic (vs. neutral non-empathic) response to negatively portrayed Ann ($M = 4.29$, $SD = 1.60$ vs. $M = 3.87$, $SD = 1.01$), $t(734) = 2.55$, $p = .033$, $d = 0.32$, $CI_{98.3\%}$ [0.02, 0.63].[22] Neither of the other two pairwise comparisons was significant (ts < 1.49, ps > .414).

[22] We used the Dunn-Bonferroni correction for all post-hoc pairwise comparisons. The confidence intervals from those comparisons correspond to the corrected α level of .017, rather than α = .05 (Dunn, 1961).

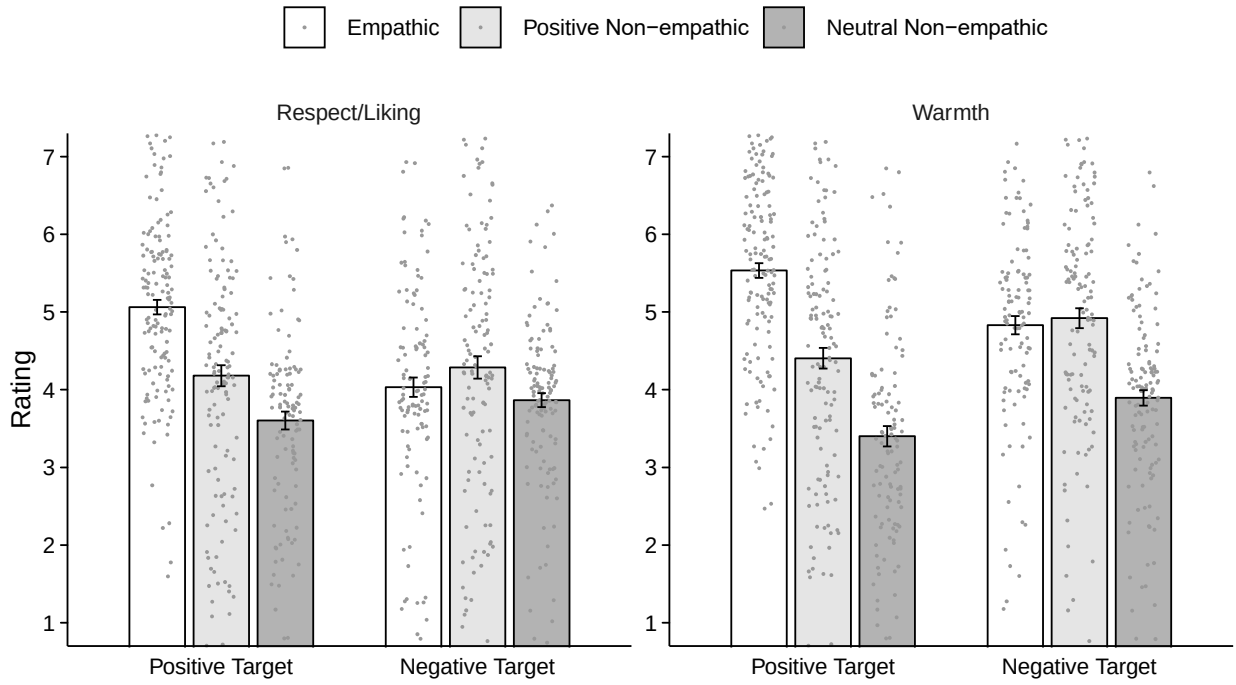

Figure 2.8. Ratings of Beth on respect/liking and warmth by response type and target valence in Experiment 4. Error bars depict ±1 standard errors; dots depict jittered individual data points.

Warmth. An identical 3 × 2 ANOVA on warmth (α = .92) revealed that, as in the previous experiments, there was a response type main effect, $F(2, 734) = 86.48$, $p < .001$, $\eta_p^2 = .19$, $CI_{90\%}$ [.15, .23]. The target valence main effect was not significant, $F(1, 734) = 1.12$, $p = .291$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .01]. More importantly, there was a significant response type × target valence interaction, $F(2, 734) = 17.63$, $p < .001$, $\eta_p^2 = .05$, $CI_{90\%}$ [.02, .07]. Planned contrasts in the positive target condition indicated that participants rated Beth as warmer when she gave an empathic (vs. positive non-empathic) response ($M = 5.53$, $SD = 1.11$ vs. $M = 4.41$, $SD = 1.51$), $t(734) = 7.09$, $p < .001$, $d = 0.87$, $CI_{95\%}$ [0.60, 1.13], and when she gave an empathic (vs. neutral non-empathic) response ($M = 3.40$, $SD = 1.37$), $t(734) = 12.86$, $p < .001$, $d = 1.64$, $CI_{95\%}$ [1.32, 1.95]. Participants also rated Beth as warmer when she gave a positive non-empathic (vs. neutral non-empathic) response to positively portrayed Ann, $t(734) = 5.95$, $p < .001$, $d = 0.77$, $CI_{95\%}$ [0.50, 1.04] (Figure 2.8, right panel).

Though not planned, we also explored whether ratings of warmth differed among the three response types in the negative target condition. Post-hoc pairwise comparisons indicated that participants rated Beth as warmer when she gave an empathic (vs. neutral non-empathic) response (M = 4.83, SD = 1.23 vs. M = 3.89, SD = 1.14), $t(734) = 5.52$, $p < .001$, $d = 0.72$, $CI_{98.3\%}$ [0.38, 1.05], and when she gave a positive non-empathic (M = 4.92, SD = 1.43) versus neutral non-empathic response, $t(734) = 6.24$, $p < .001$, $d = 0.79$, $CI_{98.3\%}$ [0.46, 1.11]. The difference in warmth ratings between the two positive response conditions was not significant, $t(734) = 0.53$, $p > .999$, $d = 0.07$, $CI_{98.3\%}$ [-0.24, 0.38].

Latent moderated mediation analyses. We conducted latent moderated mediation analyses to test whether the response type × target valence interactions on evaluations of Beth were mediated by inferences about Beth's attitudes toward Ann. We used a latent variable approach to account for the measurement error of our mediator and thereby obtain more accurate estimates of indirect effects (Ledgerwood & Shrout, 2011). In our planned analyses, the primary predictor was response type, its levels effect-coded by empathy (+2/3 = empathic, -1/3 = positive non-empathic, -1/3 = neutral non-empathic) and positivity (+1/3 = empathic, +1/3 = positive non-empathic, -2/3 = neutral non-empathic). These effect codes allowed us to test whether the effect of response empathy (i.e., empathic vs. non-empathic responses) was mediated and whether the effect of response positivity (i.e., positive vs. neutral responses) was mediated. Though not planned, we also conducted a pair of exploratory analyses in the conditions in which Beth gave a positive response (+1 = empathic, -1 = positive non-empathic), which allowed us to isolate the pattern of mediation for the effect of empathy among comparably positive responses.

In all analyses, the moderator was target valence (+1 = positive, -1 = negative); the mediator was inferences about Beth's attitudes toward Ann, modeled as a latent factor indicated

by its three items. We conducted analyses separately for respect/liking and warmth, each modeled as a latent factor indicated by its four items. Here and in subsequent experiments, we used Yzerbyt et al.'s (2018) component approach, which requires the joint significance of individual parameter estimates of an indirect effect to establish its presence (see also Muller et al., 2005). This approach can also simultaneously test for first-stage moderated mediation (i.e., interaction effect is mediated),[23] second-stage moderated mediation (i.e., mediating effect is moderated), or both (Edwards & Lambert, 2007).

A summary of the models and evidence of first-stage and second-stage moderated mediation appears in Tables 2.1A and 2.1B, and parameter estimates of individual paths appear in Tables 2.2A and 2.2B. We conducted all mediation analyses here and in subsequent experiments with the *lavaan* (Version 0.6-3; Rosseel, 2012) and *semTools* (Version 0.5-1; Jorgensen et al., 2018) packages in R (Version 3.6.0; R Core Team, 2019). All models reported below fit the data reasonably well, χ^2s(57) = 269.86–580.33, *p*s < .001, *CFI* = 0.91–0.97, *TLI* = 0.90–0.96, *RMSEA* = 0.07–0.11; details of model fit are reported in the Supplemental Materials.

Moderated mediation with response empathy. We first conducted the analysis on respect/liking using response empathy as the predictor (Model 1). The response empathy × target valence interaction significantly predicted the mediator, a_{mod} = 0.22, p = .012, and the mediator significantly predicted respect/liking, b = 0.48, p < .001, suggesting the presence of first-stage moderation. In addition, response empathy significantly predicted the mediator, a = 0.95, p < .001, and the mediator × target valence interaction significantly predicted respect/liking, b_{mod} = 0.42, p < .001, suggesting the presence of second-stage moderation. Supporting these results, the

[23] Note that the key predictor × moderator interaction on the mediator is a component of the first-stage moderated mediation (denoted as a_{mod}), which is simultaneously estimated in our structural equation models with the other paths. We present results for that path in the context of the full models below.

effect of response empathy on inferences about Beth's attitudes toward Ann, the association between inferences about Beth's attitudes toward Ann and respect/liking, and the overall indirect effect were all stronger when Ann was positively portrayed, $a_{pos} = 1.17$ vs. $a_{neg} = 0.73$, $b_{pos} = 0.90$ vs. $b_{neg} = 0.06$, $a_{pos}b_{pos} = 1.05$ vs. $a_{neg}b_{neg} = 0.04$.

We then conducted the same analysis on warmth (Model 2). The results were very similar to those for respect/liking: In addition to the effect of response empathy × target valence interaction on the mediator (a_{mod}),[24] the mediator significantly predicted warmth, $b = 0.69$, $p < .001$, suggesting the presence of first-stage moderation. Moreover, in addition to the effect of response empathy on the mediator (a), the mediator × target valence interaction significantly predicted warmth, $b_{mod} = 0.39$, $p < .001$, suggesting the presence of second-stage moderation. Supporting these results, the effect of response empathy on inferences about Beth's attitudes toward Ann, the association between inferences about Beth's attitudes toward Ann and warmth, and the overall indirect effect were all stronger when Ann was positively portrayed: $a_{pos} = 1.06$ vs. $a_{neg} = 0.66$, $b_{pos} = 1.09$ vs. $b_{neg} = 0.30$, $a_{pos}b_{pos} = 1.37$ vs. $a_{neg}b_{neg} = 0.24$.

Together, Models 1 and 2 suggested that both first-stage and second-stage moderated mediation were present when comparing the effects of empathic versus non-empathic responses: The response empathy × target valence interaction on evaluations of Beth was mediated by inferences about Beth's attitudes toward Ann; associations between the mediator and both respect/liking and warmth, in turn, were moderated by target valence.

[24] Because Models 1 and 2 have the same predictor and mediator, estimates across the two models are close to identical for a and a_{mod}. This is also the case for estimates of a and a_{mod} in Models 3 and 4 and in Models 5 and 6.

Table 2.1A.

Summary of Evidence for First-Stage Moderated Mediation in Experiment 4.

Model	Predictor	DV	First-Stage Moderated Mediation	$a_{mod}b$ [$CI_{95\%}$]	*p*
1	Empathy	Respect/liking	Yes	0.10 [0.02, 0.19]	.015
2		Warmth	Yes	0.16 [0.04, 0.30]	.014
3	Positivity	Respect/liking	No	-0.04 [-0.12, 0.04]	.358
4		Warmth	No	-0.05 [-0.17, 0.06]	.352
5	Empathic vs. positive	Respect/liking	Yes	0.09 [0.05, 0.15]	< .001
6	non-empathic	Warmth	Yes	0.14 [0.07, 0.22]	< .001

Note. Evidence of first-stage moderated mediation was determined by the joint significance of both a_{mod} and *b*.

Table 2.1B.

Summary of Evidence for Second-Stage Moderated Mediation in Experiment 4.

Model	Predictor	DV	Second-Stage Moderated Mediation	ab_{mod} [$CI_{95\%}$]	*p*
1	Empathy	Respect/liking	Yes	0.40 [0.28, 0.52]	< .001
2		Warmth	Yes	0.40 [0.28, 0.53]	< .001
3	Positivity	Respect/liking	Yes	0.95 [0.74, 1.18]	< .001
4		Warmth	Yes	1.02 [0.79, 1.26]	< .001
5	Empathic vs. positive	Respect/liking	Yes	0.08 [0.03, 0.13]	.002
6	non-empathic	Warmth	Yes	0.08 [0.03, 0.13]	.002

Note. Evidence of second-stage moderated mediation was determined by the joint significance of both *a* and b_{mod}.

Table 2.2A.

Parameter Estimates and 95% Confidence Intervals from the Latent Moderated Mediation Models 1–3 in Experiment 4.

Parameter	Model 1	Model 2	Model 3
a	0.95 [0.77, 1.13]	0.94 [0.76, 1.12]	1.78 [1.56, 2.00]
a_{pos}	1.17 [0.92, 1.41]	1.15 [0.91, 1.40]	1.70 [1.41, 1.98]
a_{neg}	0.73 [0.48, 0.99]	0.72 [0.47, 0.97]	1.86 [1.58, 2.15]
a_{mod}	0.22 [0.05, 0.39]	0.22 [0.05, 0.38]	-0.08 [-0.26, 0.09]
b	0.48 [0.38, 0.58]	0.76 [0.64, 0.88]	0.44 [0.34, 0.54]
b_{pos}	0.90 [0.75, 1.04]	1.18 [1.01, 1.35]	0.97 [0.82, 1.13]
b_{neg}	0.06 [-0.06, 0.19]	0.34 [0.20, 0.47]	-0.10 [-0.23, 0.03]
b_{mod}	0.42 [0.33, 0.51]	0.42 [0.32, 0.52]	0.54 [0.44, 0.64]
c	0.44 [0.26, 0.62]	0.88 [0.67, 1.09]	0.67 [0.48, 0.85]
c'	-0.02 [-0.20, 0.17]	0.17 [-0.03, 0.36]	-0.11 [-0.35, 0.13]
$a_{pos}b_{pos}$	1.05 [0.78, 1.34]	1.37 [1.03, 1.73]	1.65 [1.29, 2.05]
$a_{neg}b_{neg}$	0.04 [-0.04, 0.14]	0.24 [0.13, 0.38]	-0.18 [-0.43, 0.06]

Table 2.2B.

Parameter Estimates and 95% Confidence Intervals from the Latent Moderated Mediation Models 4–6 in Experiment 4.

Parameter	Model 4	Model 5	Model 6
a	1.76 [1.54, 1.98]	0.18 [0.08, 0.29]	0.18 [0.08, 0.29]
a_{pos}	1.68 [1.40, 1.96]	0.43 [0.28, 0.58]	0.43 [0.28, 0.58]
a_{neg}	1.85 [1.57, 2.13]	-0.06 [-0.21, 0.09]	-0.06 [-0.21, 0.09]
a_{mod}	-0.08 [-0.26, 0.09]	0.25 [0.14, 0.35]	0.25 [0.14, 0.35]
b	0.65 [0.53, 0.77]	0.37 [0.25, 0.49]	0.57 [0.43, 0.71]
b_{pos}	1.23 [1.05, 1.41]	0.78 [0.60, 0.96]	0.99 [0.79, 1.20]
b_{neg}	0.07 [-0.06, 0.21]	-0.04 [-0.20, 0.12]	0.14 [-0.02, 0.31]
b_{mod}	0.58 [0.47, 0.68]	0.41 [0.29, 0.53]	0.42 [0.30, 0.55]
c	1.28 [1.06, 1.49]	0.06 [-0.04, 0.16]	0.15 [0.05, 0.26]
c'	0.13 [-0.12, 0.38]	-0.01 [-0.10, 0.09]	0.05 [-0.05, 0.15]
$a_{pos}b_{pos}$	2.06 [1.62, 2.54]	0.34 [0.21, 0.49]	0.43 [0.27, 0.61]
$a_{neg}b_{neg}$	0.14 [-0.12, 0.39]	0.00 [-0.01, 0.02]	-0.01 [-0.04, 0.01]

Moderated mediation with response positivity. We next conducted our planned moderated mediation analyses using response positivity as the predictor. Unlike the results with response empathy (Models 1–2), there was evidence for second-stage moderation but not first-stage moderation for both respect/liking (Model 3) and warmth (Model 4). Supporting these results, the effect of response positivity on inferences about Beth's attitudes toward Ann was similar across target valence, but the association between the mediator and both respect/liking and warmth were stronger when Ann was positively portrayed, and the overall indirect effects were also stronger when Ann was positively portrayed (see the Supplemental Materials for detailed descriptions of Models 3 and 4). That is, inferences about Beth's attitudes toward Ann mediated the response positivity × target valence interaction on evaluations of Beth, but such inferences were predicted only by response positivity and did not differ by target valence.

Exploratory analysis. Lastly, we explored within the empathic and positive non-empathic response conditions whether the response type × target valence interaction on evaluations of Beth was mediated. Similar to results from Models 1–2, there was again evidence for both first-stage and second-stage moderation for both respect/liking (Model 5) and warmth (Model 6). Supporting these results, the effects of the empathic (vs. positive non-empathic) response on inferences about Beth's attitudes toward Ann, the associations between the mediator and evaluations of Beth, and the overall indirect effects were all stronger when Ann was positively portrayed (see the Supplemental Materials for details about Models 5 and 6). That is, even comparing only empathic versus positive non-empathic responses, inferences about Beth's attitudes toward Ann mediated the response type × target valence interaction on evaluations of Beth, and such inferences were predicted by the response type × target valence interaction.

Discussion

Experiment 4 served two purposes. First, to determine whether the response type × target valence interaction on evaluations of empathizers was driven by the positivity of the empathic response, we added a condition in which the responder gave a positive but non-empathic response. We found that positivity contributed to, but did not fully account for, the effects of empathy: When the target was positively portrayed, the empathic response elicited more respect/liking and higher ratings of warmth than did the comparably positive but non-empathic response, and both responses elicited more respect/liking and warmth than did the neutral non-empathic response. When the target was negatively portrayed, all responses elicited comparable respect/liking. The empathic and positive non-empathic responses, however, elicited comparable ratings of warmth that were higher than those elicited by the neutral, non-empathic response, suggesting that the effect of response type on warmth in the negative target conditions might be due to response positivity.

We also examined, in a series of latent moderated mediation analyses, whether inferences about the responder's attitudes toward the target mediated the response type × target valence interaction on evaluations of the responder. Although the strength and pattern of moderated mediation differed somewhat by model, evidence of moderated mediation emerged in all models. Overall, the presence of second-stage moderated mediation across all models indicates that inferences about Beth's attitudes toward Ann were more strongly associated with evaluations of Beth when Ann was positively portrayed. The presence of first-stage moderated mediation in all models with empathy contrasts (Models 1–2 and 5–6) indicates that such inferences were moderated by target valence: Whether Beth responded empathically had a stronger effect on inferences about Beth's attitude toward Ann when Ann was positively portrayed; these inferences, in turn, were associated with evaluations of Beth. The absence of first-stage

moderated mediation in models with the positivity contrast (Models 3–4) further suggests that the effect of Beth's response positivity on inferences about Beth's attitude toward Ann was not moderated by how Ann was portrayed.

In sum, Experiment 4 replicated the results of Experiments 1–2: The effects of the empathic (vs. non-empathic) response on evaluations of the responder depended on target valence. This pattern of results was partly due to the positivity of the empathic response, but the empathic response had distinct effects that differed from a comparably positive but non-empathic response. We also found evidence consistent with the possibility that participants drew inferences about the responder's attitude toward the target and used this information to form their own evaluations of the responder, though other models and/or mediators might also be consistent with the data. We return to this point in the General Discussion.

Experiment 5

In Experiments 1–4, we operationalized empathic versus non-empathic responding as the presence versus absence of empathy. Non-empathic responses, however, can take another form: The responder can actively withhold empathy from the target. In circumstances where someone responds to a generally disliked target, actively withholding empathy (e.g., expressing condemnation) unambiguously reveals how the responder views the target, which should afford evaluations of the responder. In Experiment 5, we tested this possibility by using a scenario in which a responder gave an empathic versus condemning response to a negatively portrayed target. As before, we examined the effect of response type on evaluations of the responder. We also assessed whether this effect is mediated by inferences about the responder's attitudes toward the target; we present these results, along with results from the same analysis in Experiments 6–7, in a later section (see *Mediational Evidence in Experiments 5–7*).

Method

Participants. Participants were 504 MTurk workers (50% female, 41% male, 9% no gender information; M_{age} = 39.0, SD_{age} = 11.9). We excluded n = 52 from data analyses based on our *a priori* exclusion criteria. The final sample was N = 452.

Materials and procedure. Participants were randomly assigned to one of two between-subjects conditions based on response type (empathic vs. condemning). All participants saw the same target information from the negative target conditions in the previous experiments and completed the manipulation check on target valence. Participants then read the same instructions and Ann's statement from Experiments 1 and 4, followed by Beth's response. Beth's response in the *empathic response* condition was the same as that in Experiments 1 and 4 ("I feel for you—I can really put myself in your shoes in this situation."). In the *condemning response* condition, Beth said, "To be honest, it sounds to me like you're getting what you deserve."

We collected the same dependent measures and attention checks as in previous experiments (with updated options for the attention check on Beth's response to Ann). We used the same manipulation checks from Experiments 2–4 and the same measure of inferences about Beth's attitudes toward Ann from Experiment 4.

Results

Respect/liking. An independent samples t-test on respect/liking (α = .96) indicated that participants respected/liked Beth more when she gave a condemning (vs. empathic) response to Ann (M = 4.81, SD = 1.63 vs. M = 3.58, SD = 1.74), $t(445) = 7.72$, $p < .001$, d = 0.73, $CI_{95\%}$ [0.54, 0.92] (Figure 2.9, left panel).

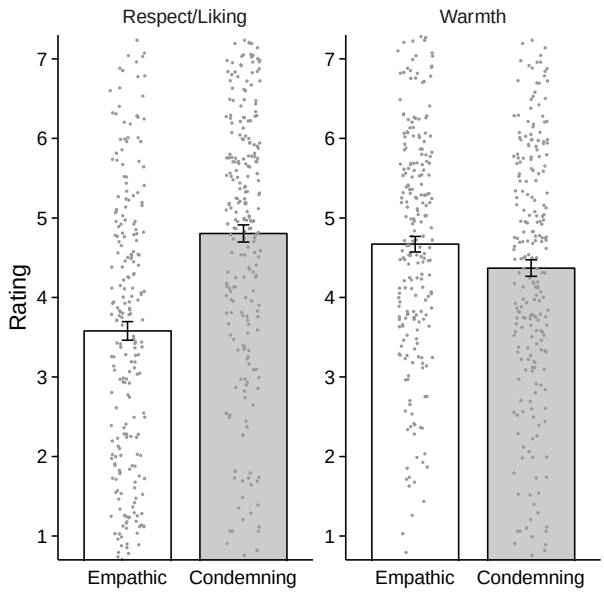

Figure 2.9. Ratings of Beth on respect/liking and warmth by response type in Experiment 5. Error bars depict ±1 standard errors; dots depict jittered individual data points.

Warmth. An independent samples *t*-test on warmth (α = .89) indicated that participants rated Beth as warmer when she gave an empathic (vs. condemning) response to Ann (M = 4.67, SD = 1.44 vs. M = 4.37, SD = 1.57), $t(449) = 2.12$, $p = .035$, $d = 0.20$, $CI_{95\%}$ [0.01, 0.38] (Figure 2.9, right panel).

Discussion

Experiment 5 examined the effect of a condemning (vs. empathic) response to a negatively portrayed target on evaluations of the responder. Whereas the condemning (vs. empathic) response increased respect/liking, it reduced the responder's warmth. Insofar as condemnation and empathy reflect negative and positive views of the negatively portrayed target, respectively, results for respect/liking align with a balanced affective triad in which participants preferred a responder who condemned (vs. empathized with) a disliked target. Results for warmth, however, are inconsistent with a balanced triad: Participants rated the responder as less warm when she condemned (vs. empathized with) a disliked target.

One potential explanation for the dissociation between respect/liking and warmth here is that they reflect judgments of the responder's morality and sociability, respectively. A growing literature indicates that morality and sociability play different roles in impression formation (e.g., Cottrell et al., 2007; Goodwin et al., 2014; Landy et al., 2016). Accordingly, it is possible that participants drew from their views on Beth's morality (e.g., her values) in evaluating if they respected/liked her, and they drew from their views on Beth's sociability in evaluating if they considered her warm. Another potential explanation for the dissociation is that participants might have relied on their evaluation of the empathic response itself, rather than that of the responder, in rating the responder's warmth. This explanation draws from research on the "act–person dissociation" in moral judgment, in which evaluation of a person can differ in valence from evaluation of an act performed by that person (Tannenbaum et al., 2011; Uhlmann et al., 2015). We revisit both explanations in the General Discussion.

Experiment 6

Experiments 1–5 used scenarios in which both the responder and the target were women. Might empathy between men be evaluated differently? On one hand, neither our target valence manipulation nor our response type manipulation was gender-specific, and we expect similar processes to operate in evaluating male versus female empathizers. On the other hand, prescriptive gender stereotypes suggest that women are expected to be warmer, kinder, friendlier, and more emotional than men, whereas men are expected to be more principled and aggressive than women (Prentice & Carranza, 2002). These gender stereotypes might, in turn, serve as standards of comparison when people evaluate male versus female empathizers and thereby produce gender differences in such evaluations. Therefore, we conducted Experiment 6 to test if the effects observed in Experiment 5 are moderated by the gender of the characters.

Method

Participants. We publicly pre-registered our analysis plan on AsPredicted (https://aspredicted.org/blind.php?x=89g5fh). Participants were 566 MTurk workers (48% female, 52% male; M_{age} = 36.3, SD_{age} = 10.9). We excluded n = 162 from data analyses based on our *a priori* exclusion criteria. The final sample was N = 404.

Materials and procedure. Participants were randomly assigned to one of the 2 (response type: empathic vs. condemning) × 2 (character gender: female vs. male) between-subjects conditions. This experiment was almost identical to Experiment 5, except for the following changes. In the male character conditions, we changed the target's name to Adam and the responder's name to Bill. We also modified the wording of the target's dialogue so that it sounded natural in both male-male and female-female interactions:

> Adam/Ann: "Work has been killing me lately. I'm organizing a rally in City Park, and we're expecting a huge turnout. The city council has been giving me a hard time with the permits. They were supposed to come through weeks ago, but they keep getting delayed. The stress is really getting to me. I feel like I haven't slept in days."

Beth/Bill then gave the same empathic response ("I feel for you—I can really put myself in your shoes in this situation") or condemning response ("To be honest, it sounds to me like you're getting what you deserve") from Experiment 5.

Results

Respect/liking. A 2 (response type: empathic vs. condemning) × 2 (character gender: female vs. male) between-subjects ANOVA on respect/liking (α = .97) revealed that participants respected/liked the condemning (vs. empathic) responder more, $F(1, 400) = 47.10$, $p < .001$, $\eta_p^2 = .11$, $CI_{90\%}$ [.06, .15]. The character gender main effect was not significant, $F(1, 400) = 0.21$, $p = .643$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .01]. The response type × character gender interaction was not

significant either, $F(1, 400) = 2.62$, $p = .106$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .03], suggesting that character gender did not moderate the effect of response type on respect/liking (Figure 2.10, left panel).

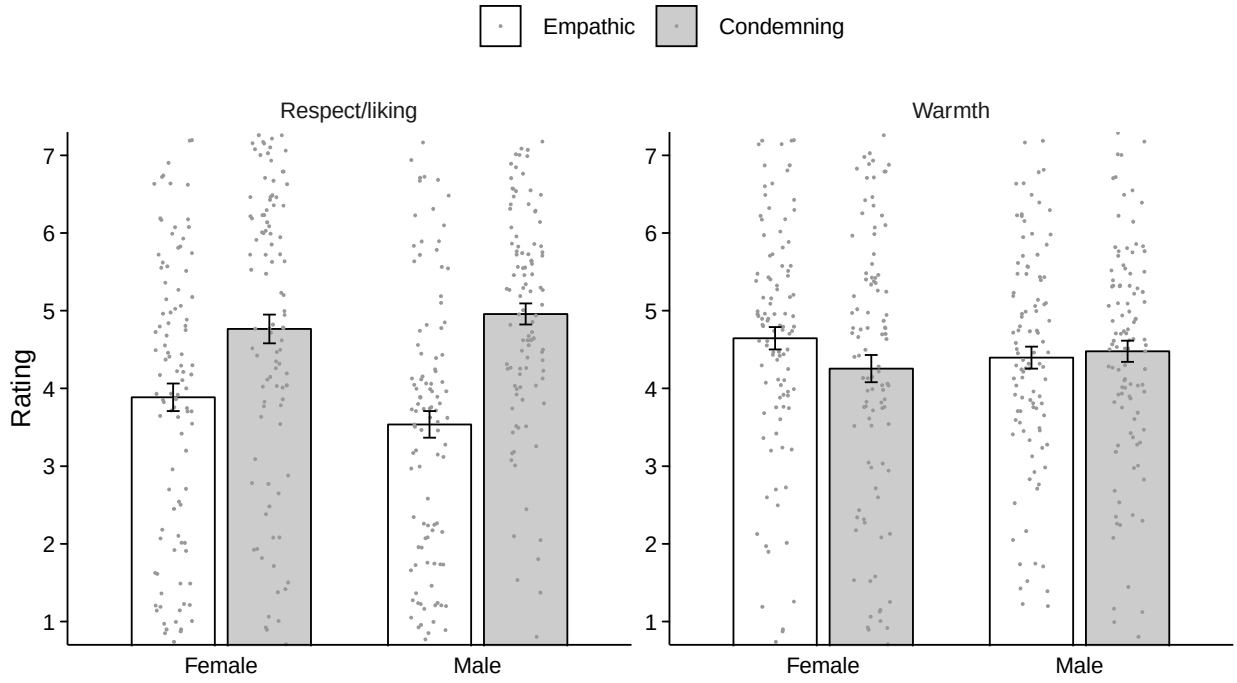

Figure 2.10. Ratings of the responder on respect/liking and warmth by response type and character gender in Experiment 6. Error bars depict ±1 standard errors; dots depict jittered individual data points.

Warmth. An identical ANOVA on warmth ($\alpha = .88$) revealed no significant effects of response type, $F(1, 400) = 0.01$, $p = .929$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .00], character gender, $F(1, 400) = 1.06$, $p = .304$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .02], or response type × character gender interaction, $F(1, 400) = 2.49$, $p = .116$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .03]. Although character gender did not moderate the effect of response type on warmth, we unexpectedly did not observe a response type main effect. An exploratory analysis in the female character condition (akin to Experiment 5) revealed that participants rated the responder as less warm when she gave a condemning (vs. empathic) response ($M = 4.26$, $SD = 1.74$ vs. $M = 4.65$, $SD = 1.45$), but this effect was not significant, $F(1, 400) = 3.36$, $p = .067$, $\eta_p^2 = .01$, $CI_{90\%}$ [.00, .03]. There was no effect of response type on warmth

in the male character condition (M = 4.48, SD = 1.40 vs. M = 4.40, SD = 1.43), $F(1, 400) = 0.15$, $p = .698$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .01] (Figure 2.10, right panel).

Discussion

Experiment 6 examined the effect of a condemning (vs. empathic) response to a negatively portrayed target on evaluations of the responder, and whether this effect was moderated by character gender. As in Experiment 5, the condemning (vs. empathic) response increased respect/liking for the responder; however, unlike Experiment 5, the condemning response did not reduce warmth toward the responder. Character gender did not moderate the effect of the condemning (vs. empathic) response on evaluations of the responder, suggesting that the moderating effect of character gender is absent or too small to be detected in our sample.

Experiment 7

Experiments 5–6 provided consistent evidence that actively condemning a negatively portrayed target increased respect/liking for the responder and provided mixed evidence that the same condemning response might decrease the responder's warmth. Experiment 7 aimed to replicate these effects and to determine whether these effects could be reversed. That is, are there circumstances where empathy with a negative target increases respect/liking for the responder? In our previous experiments, the target disclosed an experience that was directly tied to the target valence manipulation (i.e., feeling stressed *because of* her job). In Experiment 7, we included conditions in which the disclosed experience was unrelated to the source of target valence (i.e., stress from cancer treatment). If the effects of response type on evaluations of the responder from Experiments 5–6 require a direct link between the disclosed experience and target valence, then removing that link should produce different effects.

Method

Participants. We publicly pre-registered our analysis plan on AsPredicted (http://aspredicted.org/blind.php?x=c8i59d). Participants were 573 MTurk workers (52% female, 48% male; M_{age} = 36.9, SD_{age} = 11.8). We excluded n = 105 from data analyses based on our *a priori* exclusion criteria. The final sample was N = 468.

Materials and procedure. Participants were randomly assigned to one of the 2 (response type: empathic vs. condemning) × 2 (disclosed experience: job stress vs. cancer stress) between-subjects conditions. In the *job stress* condition, the procedure was the same as in Experiment 5: Participants learned that Ann was experiencing stress from work (Ann gave the same statement from Experiments 1 and 4). In the *cancer stress* condition, participants learned that Ann was experiencing stress from cancer treatment. The statement from Ann appears below; wording differences between the conditions are enclosed in square brackets:

> "I'm feeling really stressed. [I was recently diagnosed with cancer / I'm organizing an event], and my [doctors are expecting a long treatment / my team is expecting a large attendance]. I've been having trouble with the logistics of it, and the [starting date of my cancer treatment / date of the event] was delayed because the [chemotherapy medications I need are low in stock / we did not hear back from the city council in time]. The stress is overwhelming and has affected my sleep, and I've been feeling awful because of it."

Ann's statement in the two conditions closely parallel each other, in that she used the same affective expressions ("I'm feeling really stressed," "I've been having trouble…," "The stress is overwhelming and has affected my sleep, and I've been feeling awful because of it"), but the source of her stress was either related or unrelated to her job at Aryan Nations.

We collected the same set of measures and attention checks as before. We added exploratory measures that assessed whether participants thought what Ann experienced was due to the nature of her job (1 = *not at all*, 7 = *very much*) and whether Ann's circumstances were

within her control (1 = *completely within her control*, 7 = *completely out of her control*).[25] All other aspects of the procedure were the same as in Experiment 5.

Results

We followed our pre-analysis plan for all planned analyses reported below. We also conducted several exploratory analyses, which we report as such below.

Respect/liking. A 2 (response type: empathic vs. condemning) × 2 (disclosed experience: job stress vs. cancer stress) between-subjects ANOVA on respect/liking (α = .96) revealed that participants respected/liked Beth more when she responded to Ann's disclosure about job stress (vs. cancer stress), $F(1, 464) = 4.52$, $p = .034$, $\eta_p^2 = .01$, $CI_{90\%}$ [.0004, .03]. The response type main effect was not significant, $F(1, 464) = 2.10$, $p = .148$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .02]. More importantly, the response type × disclosed experience interaction was significant, $F(1, 464) = 40.09$, $p < .001$, $\eta_p^2 = .08$, $CI_{90\%}$ [.04, .12]. When Ann disclosed stress from her job, participants respected/liked Beth more when she gave a condemning (vs. empathic) response ($M = 4.63$, $SD = 1.62$ vs. $M = 3.88$, $SD = 1.71$), $F(1, 464) = 11.93$, $p < .001$, $\eta_p^2 = .03$, $CI_{90\%}$ [.01, .05]. This effect reversed when Ann disclosed stress from cancer treatment: Participants respected/liked Beth more when she gave an empathic (vs. condemning) response ($M = 4.53$, $SD = 1.46$ vs. $M = 3.33$, $SD = 1.82$), $F(1, 464) = 30.25$, $p < .001$, $\eta_p^2 = .06$, $CI_{90\%}$ [.03, .10] (Figure 2.11, left panel).

Warmth. An identical 2 × 2 ANOVA on warmth (α = .91) revealed that participants rated Beth as warmer when she responded to Ann's disclosure about job stress (vs. cancer stress), $F(1, 464) = 8.58$, $p = .004$, $\eta_p^2 = .02$, $CI_{90\%} = [.003, .04]$, and when she gave an empathic (vs. condemning) response, $F(1, 464) = 126.26$, $p < .001$, $\eta_p^2 = .21$, $CI_{90\%} = [.16, .27]$. More

[25] We collected additional exploratory measures on response positivity, participants' subjective ambivalence about Beth, and participants' evaluation of Ann after reading the interaction. Because those measures were not central to our research questions, we report them here for transparency but do not discuss them further.

importantly, the response type × disclosed experience interaction was significant, $F(1, 464) = 39.01$, $p < .001$, $\eta_p^2 = .08$, $CI_{90\%} = [.04, .12]$. When Ann disclosed stress from her job, participants rated Beth as warmer when she gave an empathic (vs. condemning) response ($M = 4.81$, $SD = 1.50$ vs. $M = 4.13$, $SD = 1.42$), $F(1, 464) = 12.45$, $p < .001$, $\eta_p^2 = .03$, $CI_{90\%} = [.01, .05]$. This effect was significantly larger when Ann disclosed stress from cancer ($M = 5.26$, $SD = 1.25$ vs. $M = 2.88$, $SD = 1.69$), $F(1, 464) = 152.79$, $p < .001$, $\eta_p^2 = .25$, $CI_{90\%} = [.19, .30]$ (Figure 2.11, right panel).

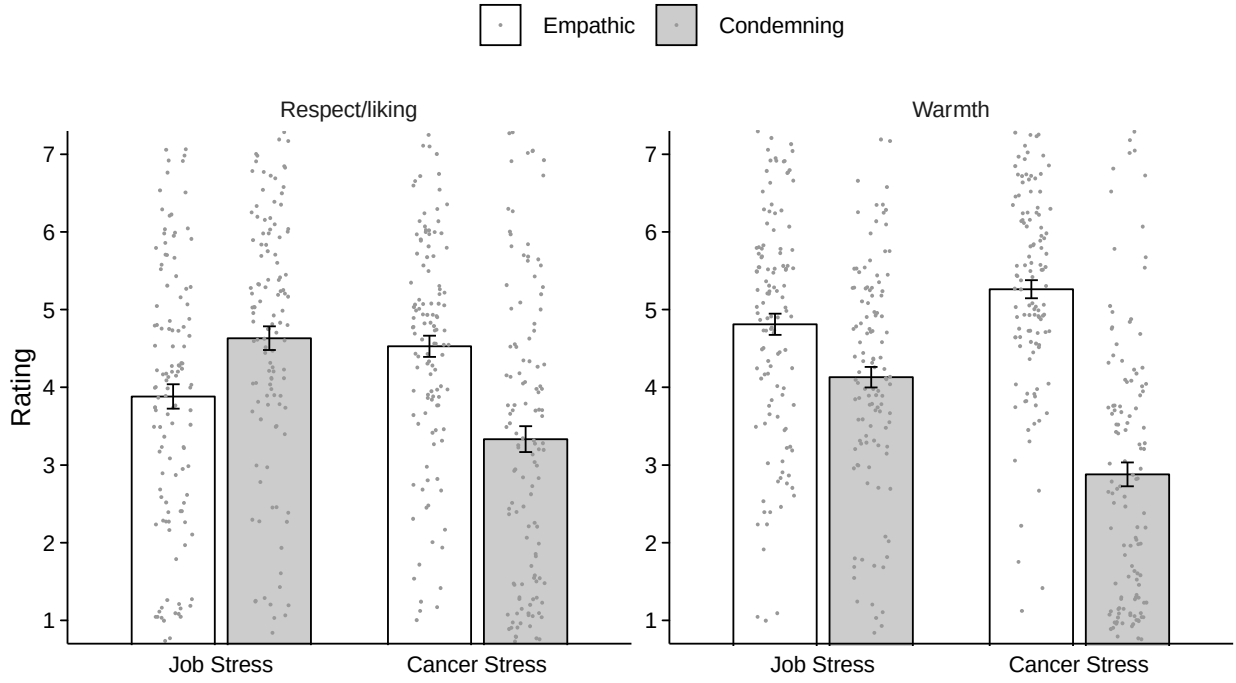

Figure 2.11. Ratings of Beth on respect/liking and warmth by response type in Experiment 7. Error bars depict ±1 standard errors; dots depict jittered individual data points.

Exploratory analysis. Lastly, we explored whether beliefs about the controllability of Ann's circumstances differed by disclosed experience. Participants thought that Ann's circumstances were more out of her control in the cancer stress (vs. job stress) condition ($M = 5.26$, $SD = 1.90$ vs. $M = 3.85$, $SD = 1.74$), $t(463) = 8.36$, $p < .001$, $d = 0.77$, $CI_{95\%}$ [0.58, 0.96].

Discussion

Experiment 7 provided additional evidence for the effects of a condemning (vs. empathic) response to a negative target on evaluations of the responder. Replicating Experiments 5–6, when a negatively portrayed target disclosed a stressful experience that was directly linked to why she was negatively evaluated (i.e., stress from her job at a white supremacist organization), participants respected/liked the responder more when she gave a condemning (vs. empathic) response. This effect reversed, however, when the target's stressful experience was unrelated to why she was negatively evaluated (i.e., stress from cancer treatment). These results suggest that respect/liking for someone who actively withholds empathy from a disliked target depends on whether the disclosed experience is directly linked to why the target is disliked. Furthermore, replicating Experiment 5 and consistent with the direction of the simple main effect in the female character condition in Experiment 6, the condemning (vs. empathic) response elicited lower ratings on the responder's warmth when the disclosed experience was directly linked to the source of target valence; this effect was even stronger when the link was absent.

One limitation of our disclosed experience manipulation is that participants in the cancer stress condition might have found the source of the target's stress jarring. Although the wording for the target's affective experience was identical in the job stress and cancer stress conditions, a life-threatening illness like cancer is arguably more stressful than work problems. We do not have direct evidence suggesting that participants viewed the target's cancer stress as more severe than job stress,[26] but the generalizability of the findings in this experiment could benefit from future research that uses alternative manipulations of the cause of stress.

[26] Our exploratory variable on ratings of Ann's affect did not significantly differ in the job stress (vs. cancer stress) conditions (M = 2.57, SD = 1.58 vs. M = 2.34, SD = 1.54), $t(466) = 1.63$, $p = .104$, $d = 0.15$, $CI_{95\%}$ [-0.03, 0.33]. It is possible, however, that participants did not believe cancer-stricken Ann felt worse, but they felt worse for her.

Mediational Evidence in Experiments 5–7

Similar to Experiment 4, in Experiments 5–7, we conducted latent mediation analyses to test whether the effects of response type on evaluations of the responder were mediated by inferences about the responder's attitudes toward the target. In these mediation models, the predictor was response type (+1/2 = condemning, -1/2 = empathic), and the mediator was inferences about the responder's attitudes toward the target, which was modeled as a latent factor indicated by its three items. As in Experiment 4, we conducted analyses separately for respect/liking and warmth, each modeled as a latent factor indicated by its four items.

We conducted simple mediation models on all data in Experiment 5, all data in Experiment 6 (due to the lack of character gender main effect on the DVs), and all data in the job stress condition in Experiment 7 (see Figure 2.12 for a model diagram). Simple mediation models were planned *a priori* in Experiments 5–6 and were exploratory in Experiment 7 (for which the planned analyses were latent moderated mediation analyses, reported below). We followed our planned analyses except for one data-dependent modeling decision: We allowed the residual covariance of the two positively-worded items in the mediator (agreement with the statements "[Responder] likes [target]" and "[Responder] feels positive toward [target]") to be freely estimated. This decision reduced model misspecification of the mediator from ignoring wording-related covariance (Marsh, 1996) and better isolated the true mediator variance, which, in turn, should provide greater power and more accurate indirect effect estimates (Gonzalez & MacKinnon, 2020). To retain local independence of the latent mediator, we also constrained the factor loadings of those two items to be equal.

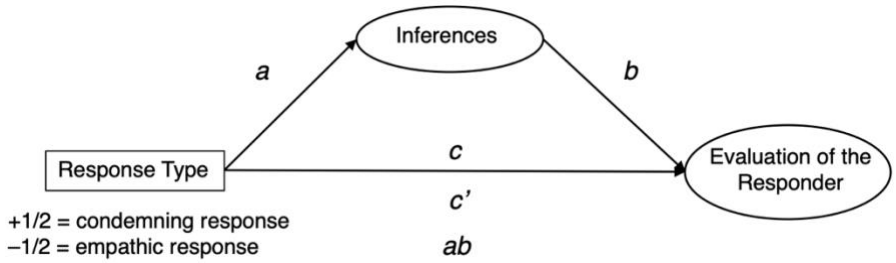

Figure 2.12. Diagram of the latent simple mediation models in Experiments 5–7. Inferences = Inferences about the responder's attitudes toward the target. Evaluation of the responder is either respect/liking or warmth. For visual clarity, measurement models are not shown.

These models fit the data reasonably well, χ^2s(18) = 86.18–233.08, *p*s < .001, *CFI* = 0.92–0.97, *TLI* = 0.87–0.95, *RMSEA* = 0.12–0.17 (see Supplemental Material for model details). Results were mixed across the three experiments: The indirect effects (*ab*) had largely overlapping 95% confidence intervals, but the point estimates differed in sign and significance for both respect/liking and warmth (Table 2.3). Integrated analyses on the pooled data from the three experiments indicated that the indirect effect was significant for respect/liking, *ab* = 0.64, *p* < .001, but not for warmth, *ab* = 0.08, *p* = .594.

Table 2.3.

Parameter Estimates, 95% Confidence Intervals, and p-Values from the Main Simple Latent Mediation Models in Experiments 5–7.

DV	Dataset	*a*	*b*	*ab*	p_{ab}
Respect/ liking	Experiment 5	-4.46 [-5.12, -3.79]	-0.30 [-0.44, -0.16]	1.34 [0.68, 2.00]	< .001
	Experiment 6	-3.25 [-3.81, -2.69]	-0.22 [-0.37, -0.08]	0.72 [0.23, 1.22]	.004
	Experiment 7	-3.36 [-4.00, -2.72]	0.09 [-0.08, 0.26]	-0.31 [-0.87, 0.26]	.289
	Pooled Data	-3.69 [-4.05, -3.33]	-0.17 [-0.26, -0.09]	0.64 [0.32, 0.96]	< .001
Warmth	Experiment 5	-4.40 [-5.05, -3.75]	-0.14 [-0.27, -0.01]	0.63 [0.05, 1.20]	.034
	Experiment 6	-3.23 [-3.79, -2.68]	-0.03 [-0.17, 0.11]	0.10 [-0.35, 0.55]	.655
	Experiment 7	-3.37 [-4.01, -2.73]	0.19 [0.02, 0.37]	-0.65 [-1.25, -0.05]	.035
	Pooled Data	-3.66 [-4.02, -3.31]	-0.02 [-0.11, 0.06]	0.08 [-0.22, 0.38]	.594

Note: All parameter estimates are in unstandardized metrics.

To assess the robustness of evidence for indirect effects from the simple mediation analyses, we compared the results with those from two alternative analytic approaches. The first alternative approach was almost identical to the main approach but ignored wording differences among the items in the mediator (i.e., the mediator items had freely estimated factor loadings and independent residual variances). The second alternative approach contained only observed (rather than latent) variables and modeled both the mediator and the DVs as composite scores. Results from these two approaches were largely consistent with our main analytic approach and revealed mixed evidence for the indirect effects (see Supplemental Material for details). We conclude that there is some evidence for mediation by inferences about the responder's attitudes toward the negative target on evaluations of the responder in Experiment 5–7. Importantly, this evidence is weak, inconsistent across experiments, and dependent on analytic approaches. Note, however, that because Experiments 5–7 only presented negative targets, the weak mediational evidence is in line with the results of Experiment 4, which found that the indirect effects were weaker in the negative (vs. positive) target conditions (see Tables 2.2A and 2.2B, $a_{neg}b_{neg}$).

Lastly, we followed our preregistered analysis plan for Experiment 7 and conducted latent moderated mediation analyses by entering disclosed experience as a moderator (+1/2 = job stress, -1/2 = cancer stress). These moderated mediation models revealed no evidence of moderated mediation for respect/liking (first-stage moderation: $a_{mod}b = 0.07$, $p = .126$; second-stage moderation: $ab_{mod} = 0.11$, $p = .526$), and weak evidence of first-stage moderated mediation for warmth (first-stage moderation: $a_{mod}b = 0.13$, $p = .034$; second-stage moderation: $ab_{mod} = 0.26$, $p = .154$). Because of the mixed results on the simple indirect effects across Experiments 5–7, we similarly conclude that the evidence of moderated mediation is weak; details of these analyses are available in the Supplemental Material.

General Discussion

Although empathy is widely studied, little is known about its effects beyond the dyadic context. The current research focuses on the extra-dyad implications of empathy for empathizers. In seven experiments, we examined how third-party observers evaluate empathizers and how target characteristics affect such evaluations. Evaluations of empathizers consistently depended on target valence. Empathizers were respected/liked more when they responded to a positively portrayed target, but not when they responded to a negatively portrayed target (Experiments 1, 2, and 4). Empathizers and non-empathizers were respected/liked comparably when a negative target shared a stressful experience (Experiments 1, 2, and 4), but empathizers were respected/liked less when a negative target shared a positive experience (i.e., success at work; Experiment 3). In addition, empathizers were rated as warmer when they responded to both positive and negative targets, but the effect was smaller for negative targets (Experiments 1–4).

We found that the effects on evaluations of empathizers were partially, but not solely attributable to the positivity of the empathic response (Experiment 4). In addition, inferences about the responder's attitudes toward the target frequently mediated the effects, though the mediational evidence was weaker when the target was negatively portrayed (Experiments 4–7). Lastly, when the responder condemned (vs. empathized with) a negatively portrayed target, they were respected/liked more but seen as less warm when the target experienced stress from working for a white supremacist organization; when the target's stressful experience was unrelated to her negative portrayal (i.e., cancer treatment), the responder was respected/liked less and also rated as less warm (Experiments 5–7).

By examining the effects of empathy beyond the dyad in which it takes place, our research expands current understanding of the social impact of empathy. Importantly, our

findings reveal that empathy not only *can* have an impact on extra-dyad observers, but also that the impact is nuanced: Observers' evaluations of empathizers are attuned to an array of target characteristics, including the target's valence, the target's experience, and the cause of the empathized experience. These findings lend credence to the idea that empathy can serve as a tool for social affiliation: By expressing empathy, empathizers signal for whom they care, which, in turn, can be used by observers to evaluate empathizers' personal character. In this spirit, the current research converges with other recent work in showing that affiliative intra-dyad phenomena can affect third-party observers' judgments and behaviors (e.g., Algoe et al., 2019; Critcher & Zayas, 2014; Kavanagh et al., 2011).

The impact of empathy on observers also poses a conundrum: People are often encouraged to empathize with disliked others, but our findings suggest that they are not always viewed favorably for doing so. Given that empathizing with liked others has evaluative benefits for empathizers, whereas empathizing with disliked others might not (or might even incur evaluative costs), these benefits and costs might, in turn, affect the reputation and social standing of empathizers. This possibility underscores the importance of considering the extra-dyadic effects of empathy, because it suggests that the effects of empathy within a dyad might not be congruent with its effects beyond the dyad. Insofar as empathy is seen as an affiliative act, it might not bridge social divides as some have claimed, because those who actually empathize across social divides might be repudiated by their own peers for doing so. Consequently, the social evaluative benefits of empathy might accrue more readily within groups than across them. Empathy might thus ironically reify the very social divides it is touted to bridge.

Dissociation Between Respect/Liking and Warmth

One intriguing finding is that respect/liking for and warmth of empathizers were dissociated when the target was negatively portrayed. To provide a cumulative picture of how respect/liking versus warmth varied by response type in the negative target conditions, we meta-analyzed the effect across all experiments using McShane and Böckenholt's (2017) single-paper meta-analysis tool. We coded the levels of response type as -1 for empathic responses and +1 for non-empathic responses (i.e., non-empathic response in Experiments 1–3 and S1, neutral non-empathic response in Experiment 4, and condemning response in Experiments 5–7) for respect/liking, and the reverse for warmth (i.e., +1 for empathic responses, -1 for non-empathic responses). The response type × dependent variable interaction was significant, $b = 0.94$, $p < .001$, $CI_{95\%}$ [0.42, 1.46], suggesting that respect/liking and warmth indeed differed by response type in the negative target conditions (see Figure 2.13 for estimates across all experiments). Because we used different manipulations of empathic versus non-empathic responses across experiments, unsurprisingly the effect sizes were heterogeneous, $I^2 = 93\%$, $CI_{95\%}$ [91%, 94%].

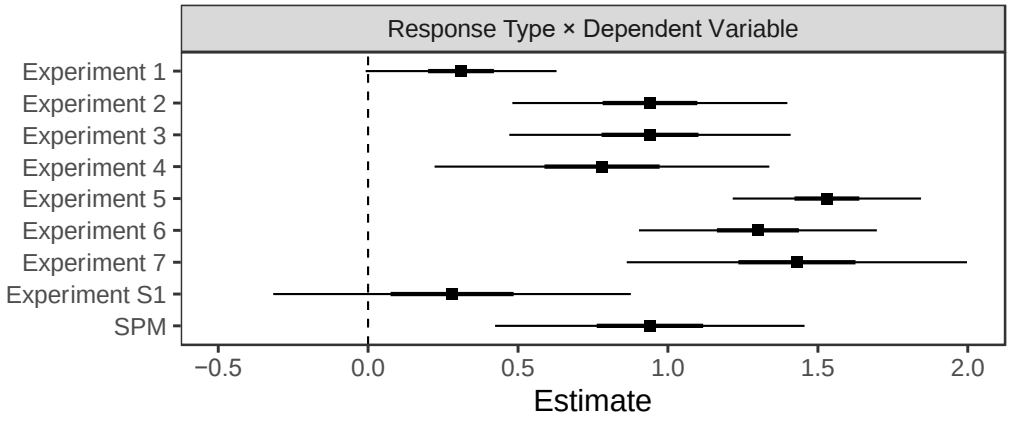

Figure 2.13. Estimates of the response type × dependent variable interaction effect in all experiments and the single-paper meta-analyzed (SPM) effect. Thick and thin lines, respectively, represent 50% and 95% confidence intervals.

Why might people evaluate empathizers of negative targets differently on respect/liking versus warmth? One perspective is that in our experimental contexts, respect/liking and warmth

reflect different evaluative dimensions. This possibility is largely consistent with the literature on the dual-dimension of social evaluation (e.g., Abele & Wojciszke, 2007; Asch, 1946; Fiske et al., 2007; Todorov et al., 2008): In our experiments, respect/liking is similar to agency/competence, and warmth is similar to warmth/communion. We note, however, that there are important points of divergence between the two dimensions we observed and the dual-dimension models. For example, our results suggest that respect and liking belong to the same dimension but that liking and warmth belong to different dimensions; in contrast, both agency–communion models and the stereotype content model view liking as reflecting warmth/communion (e.g., Asch, 1946), and that liking and respect are separate dimensions (e.g., Fiske et al., 2007; Wojciszke et al., 2009).

Another perspective on the dissociation between respect/liking and warmth in the negative target conditions is to view it through the lens of moral judgment. Although we did not assess participants' beliefs about the morality of the negative targets, it seems likely that those beliefs underlie their unfavorable views of the negative targets (i.e., participants disliked the targets because they repudiate the values those targets held). If negative views of the negative targets were moralized, then respect/liking for the empathizer might similarly result from judgments of the empathizer's morality (e.g., "Is Beth a *good* person for empathizing with a white supremacist / anti-vaxxer?"). Insofar as participants' values diverge from their inferred values of the empathizer, they should respect/like the empathizer less. In contrast, warmth might capture the empathizer's sociability (e.g., "Is Beth a *nice* person for empathizing with a white supremacist / anti-vaxxer?"). In other words, whereas warmth might reflect evaluations of what the responder is like interpersonally, respect/liking might reflect evaluations of what the responder stands for. This possibility draws from research indicating that morality and sociability represent two distinct components of person perception (Cottrell et al., 2007; Goodwin et al.,

2014; Landy et al., 2016), which might also explain why liking for the empathizer did not load onto the warmth factor: Insofar as liking reflects a global impression of the empathizer, it should be aligned with judgments of the empathizer's morality, because of the primacy of moral information in shaping global impressions (Brambilla & Leach, 2014; Landy et al., 2016).

Alternatively, drawing from the act–person dissociation in moral judgment (e.g., Pizarro & Tannenbaum, 2011; Uhlmann et al., 2015), the dissociation between respect/liking and warmth might reflect a focus on the person (i.e., empathizer) versus the act (i.e., showing empathy). For instance, Uhlmann et al. (2013) found that consequentialist actions (e.g., throwing a dying man overboard to prevent a lifeboat from sinking) can lead to positive evaluation of the action (as morally permissible) but negative evaluation of the person's character, due to the attribution that the person lacks empathy. It is possible that respect/liking reflects evaluations of the moral character of empathizers and that warmth reflects evaluations of the act of empathy (e.g., "Empathizing with a white supremacist shows care, but I don't like the person doing it.").

Limitations and Future Research Directions

Our findings indicate that expressing empathy affects observers' evaluations of empathizers. A promising direction for future research is to examine whether these evaluations have behavioral implications as well. This direction is applicable to everyday interpersonal settings, where observers might choose to affiliate with or distance themselves from empathizers. It is also applicable to more visible settings, where public figures might outwardly display empathy. For example, when a political candidate expresses empathy toward potential voters, under what conditions does that candidate garner more support? Past research suggests that empathy is among the most important and influential traits that voters consider in U.S. presidential and Senate elections (Hayes, 2005, 2010). Yet, our research suggests that voters'

support for empathic political candidates might be conditional on how voters view the recipients of the candidates' empathy (e.g., someone voters like vs. dislike). Furthermore, additional factors could influence behaviors toward empathizers more heavily than evaluations of empathizers. One such factor is authenticity: People might be willing to positively evaluate but not personally affiliate with or support empathizers who seem inauthentic or performative. Examining when evaluations of empathizers lead to affiliative behaviors could shed additional light on how empathy coordinates social behaviors.

We found that inferences about the responder's attitude toward the target largely mediated the effects of our experimental manipulations. We caution that this finding, although consistent with the possibility that people use their inferences to evaluate the responder, should be viewed as correlational evidence. The limitations of drawing causal inferences from cross-sectional mediation analyses are well-known (e.g., Bullock et al., 2010; Spencer et al., 2005), and it is possible that other variables could additionally or alternatively explain the effects reported here. Future research could clarify the causal role of inferences about the responder's attitudes toward the target by directly manipulating those inferences.

Across experiments, we used verbal information to manipulate expressions of empathy in a dyadic exchange. We did so not only for experimental control, but also because verbal information is an important medium through which people observe empathy (e.g., in printed media or online exchanges). As illustrated by the reactions to the *New York Times* profile we discussed at the outset, verbally conveyed empathy alone can elicit observers' evaluations. At the same time, other forms of expressing and observing empathy, particularly those involving live, in-person interactions, likely contain richer information and afford more nuanced inferences and evaluations. For example, in inferring an empathizer's attitudes toward a target from an in-

person interaction, observers might integrate the target's and the empathizer's nonverbal behaviors, such as their tone of voice, eye contact, and proximity to each other (e.g., DePaulo & Friedman, 1998; Dovidio et al., 2002). These factors might also contextualize verbal communications: For example, observers might interpret someone expressing empathy in an insincere tone as evidence that the empathizer does not actually like the target. Thus, our results might not generalize to live empathic interactions in which a wider range of variables might affect how observers evaluate empathizers.

Finally, we focused on observers' impressions of empathizers in cases where observers had no existing relationship with the empathizer or the target. What would happen if the observer does? For example, how would someone evaluate their own parent for empathizing with a disliked person? Unlike the impression formation context we examined, established relationships between observers and empathizers afford observers knowledge about and attitudes toward empathizers that observers might be motivated to preserve (Festinger, 1957; Kunda, 1990). One possibility in the context of established relationships is that observers who already view the empathizer positively might be motivated to "explain away" empathy with a negative target (e.g., "She might empathize with white supremacists, but she's still a good person") or reinterpret empathy in a positive way (e.g., "She empathizes with white supremacists because she sees the good in everyone"). We consider this possibility a promising direction for future research.

Conclusion

Empathy is often considered a virtue, yet people who display it might not always be viewed positively. The present work indicates that third-party observers' evaluations of empathizers crucially depend on the target of empathy. More broadly, our findings underscore the extra-dyadic effects that empathy has: Empathy connects people, but the connections it

enables have evaluative consequences. Understanding how people view empathy and empathizers promises a deeper understanding of how empathy functions in social contexts.

Supplemental Material

Experiment 1

Participants and Power Analysis

Amazon's Mechanical Turk (MTurk) workers (N = 464) completed the experiment online for modest remuneration. Because this was the first experiment in this line of research, we did not have an effect size estimate and thus set a target sample size that would provide 80% power (α = .05) to detect a small effect (η_p^2 = .02) in a 2 × 2 between-subjects design. A power analysis[27] suggested a minimum sample size of N = 387. We anticipated an exclusion rate of 10% and decided to collect data from N = 450.[28] We decided *a priori* on the following exclusion criteria: failing the attention check on Beth's response to Ann, failing the attention check on Ann's employer, and giving identical responses across all dependent variables (because we had reverse-coded items). Upon concluding data collection but prior to analysis, we decided to retain data from participants who gave identical neutral responses (i.e., 4 on a 7-point scale) across the dependent variables; we reasoned that one could plausibly feel neutral on all items (this modification did not change our results). The numbers of participants who met each exclusion criterion were n = 89, n = 52, and n = 7, respectively. The final sample was N = 336 (some participants met more than one criterion).[29]

Manipulation Checks

The manipulations of target valence and response type were both successful: Participants in the positive target (vs. negative target) conditions viewed Ann more positively (M = 5.83, *SD*

[27] Across experiments, we conducted all power analyses with G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), unless otherwise noted.
[28] The number of participants initially included in this and subsequent samples was slightly higher than our target sample size because our data collection platform counted the number of people who proceeded to the last page of our experiment rather than the number of people who completed all survey questions.
[29] Due to a programming oversight, we did not collect information on participant gender and age in Experiments 1 and S1. We report participant gender and age for all other experiments.

= 1.00 vs. M = 1.97, SD = 1.58), $t(255) = 26.37$, $p < .001$, $d = 2.97$, $CI_{95\%}$ [2.66, 3.29], and participants in the empathic (vs. non-empathic) response conditions thought that Beth empathized with Ann to a greater extent (M = 5.83, SD = 1.15 vs. M = 4.03, SD = 1.46), $t(263) = 12.26$, $p < .001$, $d = 1.40$, $CI_{95\%}$ [1.16, 1.64].

Experiment S1

Experiment S1 was a conceptual replication of Experiment 1. We used the same experimental design but extended the dialogue to clarify that Beth and Ann did not know each other beforehand. The dialogue also included Beth's confirmation that she knew where Ann worked, thus removing ambiguity about whether Beth understood the mission of Ann's employer. Despite these changes, we expected to replicate the results of Experiment 1.

Method

Participants and power analysis. MTurk workers (N = 472) participated online for modest remuneration. A power analysis indicated that a sample size of N = 296 affords 80% power ($\alpha = .05$) to detect an effect comparable in size to the key interaction effects in Experiment 1 (around $\eta_p^2 = .026$). Given the 28% exclusion rate in Experiment 1 and the need to exclude participants who had completed Experiment 1, we decided to match the sample size of Experiment 1 and collect data from N = 450. As in Experiment 1, we decided *a priori* to exclude data from participants who failed the attention check on Beth's response to Ann, failed the attention check on Ann's employer, or gave identical non-neutral responses (i.e., other than 4 on 7-point scales) across all dependent variables. We also decided *a priori* to exclude data from participants who indicated that they had completed Experiment 1.[30] The numbers of participants

[30] In subsequent experiments, only MTurk workers who had not already participated in a study in this line of research were eligible to participate, so this data exclusion criterion was not used in Experiments 2–7.

who met each criterion were $n = 56$, $n = 43$, $n = 5$, and $n = 48$, respectively. The final sample was $N = 373$ (some participants met more than one exclusion criterion).

Materials and procedure. This experiment was identical to Experiment 1, except for the differences reported here. Participants (a) learned that Beth and Ann were meeting for the first time at a neighborhood dog park and (b) read a more extensive dialogue, during which Ann revealed the organization she worked for (text for the positive target conditions appears below; in the negative target conditions, the organization name was replaced with "Aryan Nations"):

> Beth: "I don't think I've ever seen you around here before. Are you new to the neighborhood?"
>
> Ann: "Yes, I just moved here. My name is Ann. Nice to meet you."
>
> Beth: "Nice to meet you! I'm Beth. How are you doing?"
>
> Ann: "Well...not so great, to be honest."
>
> Beth: "How come?"
>
> Ann: "I'm feeling really stressed. I work for this organization called St. Jude Children's Research Hospital. Are you familiar with it?"
>
> Beth: "Yes, I've heard of it."
>
> Ann: "So yeah, I do event planning and outreach for St. Jude Children's Research Hospital, and I'm organizing an event for them. My team is expecting a large attendance, but I've been having a lot of trouble with the logistics of it, and the date of the event was recently delayed because we did not hear back from the city council in time. The stress is overwhelming and has affected my sleep, and I've been feeling awful because of it."

The options for the attention check on Beth's response to Ann included Beth's full responses ("I feel for you—I can really put myself in your shoes in this situation. When is the event taking place?", "Okay, I see. When is the event taking place?", "I don't understand your situation. When is the event taking place?", and *none of the above*).

Results

Manipulation checks. Both manipulations were again successful: Participants in the positive (vs. negative) target conditions viewed Ann more positively (M = 5.82, SD = 1.08 vs. M = 1.82, SD = 1.46), $t(346) = 30.17$, $p < .001$, $d = 3.11$, $CI_{95\%}$ [2.81, 3.42], and participants in the empathic (vs. non-empathic) response conditions indicated that Beth empathized with Ann to a greater extent (M = 5.77, SD = 1.25 vs. M = 4.35, SD = 1.42), $t(365) = 10.23$, $p < .001$, $d = 1.06$, $CI_{95\%}$ [0.84, 1.28].

Factor analysis. To confirm the factor structure from Experiment 1, we conducted a confirmatory factor analysis (CFA) in R using the *lavaan* package (Rosseel, 2012). Drawing from the EFA solution in Experiment 1, we specified a model with two latent factors; four items (*like*, *respect*, *trust*, and *friends*) loaded onto the first factor (*respect/liking*), and the other four items (*understanding*, *kind*, *cold* [reverse-coded], and *caring*) loaded onto the second factor (*warmth*). Because factor loadings of all items on their non-primary factors were low in the EFA solution in Experiment 1, we specified no cross-loadings in the CFA. This two-factor model fit the data well, $\chi^2(19) = 59.96$, $p < .001$, $RMSEA = 0.08$, $CFI = 0.99$, $TLI = 0.98$, with all factor loadings higher than $\lambda = .60$. The two-factor model also fit the data better than a one-factor model in which all items loaded onto a single factor, $\Delta\chi^2(1) = 184.57$, $p < .001$. Thus, we confirmed the factor structure from Experiment 1. As in Experiment 1, we calculated the mean ratings of items for respect/liking ($\alpha = .96$) and warmth ($\alpha = .90$) as composites and conducted the primary analyses on these composites.

Respect/liking. A 2 (response type) × 2 (target valence) between-subjects ANOVA on respect/liking again yielded main effects of both factors: Participants respected/liked Beth more when she gave an empathic (vs. non-empathic) response, $F(1, 369) = 7.44$, $p = .007$, $\eta_p^2 = .02$,

$CI_{90\%}$ [.003, .05], and when Ann was positively (vs. negatively) portrayed, $F(1, 369) = 123.73$, $p < .001$, $\eta_p^2 = .25$, $CI_{90\%}$ [.19, .31]. The response type × target valence interaction was marginally significant, $F(1, 369) = 3.13$, $p = .078$, $\eta_p^2 = .01$, $CI_{90\%}$ [.00, .03]. When Ann was positively portrayed, participants respected/liked Beth more when she gave an empathic (vs. non-empathic) response ($M = 5.57$, $SD = 1.13$ vs. $M = 4.94$, $SD = 0.96$), $F(1, 369) = 9.96$, $p = .002$, $\eta_p^2 = .03$, $CI_{90\%}$ [.01, .06]. When Ann was negatively portrayed, however, respect/liking for Beth did not significantly differ by response type ($M = 3.77$, $SD = 1.72$ vs. $M = 3.64$, $SD = 1.44$), $F(1, 369) = 0.47$, $p = .495$, $\eta_p^2 < .01$, $CI_{90\%}$ [.00, .01] (see Figure S2.1, left panel).

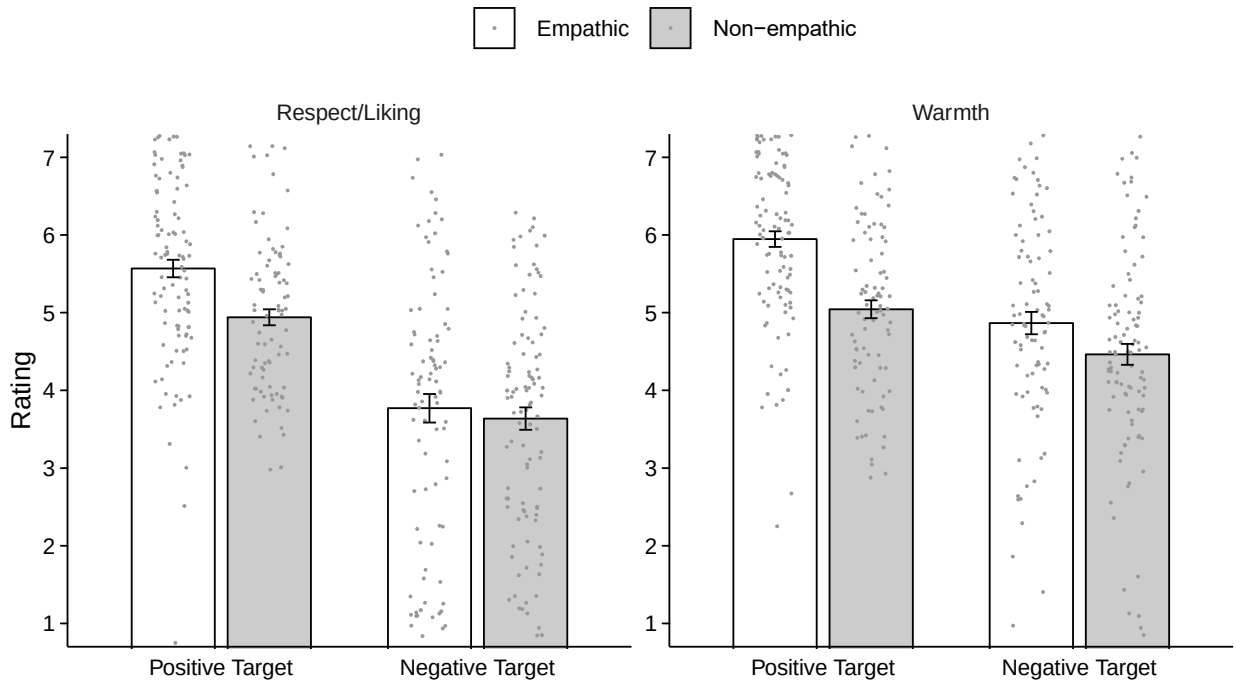

Figure S2.1. Ratings of Beth on respect/liking and warmth by response type and target valence in Experiment S1. Error bars depict ±1 standard errors; dots depict jittered individual data points.

Warmth. An identical 2 × 2 ANOVA on warmth also revealed main effects of response type and target valence: Participants rated Beth as warmer when she gave an empathic (vs. non-empathic) response, $F(1, 369) = 27.39$, $p < .001$, $\eta_p^2 = .07$, $CI_{90\%}$ [.03, .11], and when Ann was positively (vs. negatively) portrayed, $F(1, 369) = 44.32$, $p < .001$, $\eta_p^2 = .11$, $CI_{90\%}$ [.06, .16]. The

response type × target valence interaction was (barely) significant, $F(1, 369) = 4.00$, $p = .046$, $\eta_p^2 = .01$, $CI_{90\%}$ [.0001, .03]. When Ann was positively portrayed, participants rated Beth as warmer when she gave an empathic (vs. non-empathic) response ($M = 5.95$, $SD = 1.01$ vs. $M = 5.04$, $SD = 1.06$), $F(1, 369) = 25.80$, $p < .001$, $\eta_p^2 = .07$, $CI_{90\%}$ [.03, .11]. Unlike the respect/liking results, participants rated Beth as warmer when she gave an empathic (vs. non-empathic) response to negatively portrayed Ann ($M = 4.87$, $SD = 1.34$ vs. $M = 4.46$, $SD = 1.35$), though the effect was smaller, $F(1, 369) = 5.30$, $p = .022$, $\eta_p^2 = .01$, $CI_{90\%}$ [.001, .04] (see Figure S2.1, right panel).

Discussion

Experiment S1 largely replicated the results of Experiment 1. First, we confirmed the same two-dimensional structure of evaluations of the responder. Second, using an extended, less ambiguous dialogue, we again found that evaluations of empathizers depended on the target of empathy (though the response type × target valence interaction on respect/liking was marginally significant). Participants respected/liked the responder more when she responded empathically to a positively portrayed target, but not when she responded to a negatively portrayed target. Participants also rated the responder as warmer when she responded empathically, but this effect was smaller when the target was negatively portrayed. Although the effect sizes were smaller here than in Experiment 1, the overall pattern of results was largely unaffected by assumptions about Beth and Ann's relationship or Beth's knowledge about Ann's employer.

Experiment 2

Participants and Power Analysis

We publicly pre-registered our analysis plan, including power analyses, target sample size, inclusion and exclusion criteria, and planned data analyses, on AsPredicted (http://aspredicted.org/blind.php?x=fr6rn9). MTurk workers ($N = 614$, 49% female, 51% male;

M_{age} = 37.9, *SD* = 12.5) participated online for modest remuneration. We planned to power this experiment at 80% (α = .05) to detect the expected effect of target valence × response type interaction on respect/liking and warmth. Based on past experiments, we estimate the interaction effect sizes as η_p^2 = .017 (respect/liking) and .029 (warmth). The sample sizes required to detect these effect sizes are *N* = 456 and *N* = 265. We chose the more conservative *N* = 456 as the target sample size for analysis. Based on an anticipated exclusion rate of 25% estimated from previous experiments, we set a target sample size of *N* = 608. We decided *a priori* to exclude participants based on the same three exclusion criteria from Experiment 1. The numbers of participants who met each exclusion criterion were *n* = 33, *n* = 70, and *n* = 2, respectively. The final sample was *N* = 526 (some participants met more than one exclusion criterion).

Manipulation Checks

All manipulations were successful. Participants evaluated Ann more positively when she was portrayed as pro- versus anti-vaccination (*M* = 5.66, *SD* = 1.19 vs. *M* = 2.80, *SD* = 1.74), *t*(442) = 21.94, *p* < .001, *d* = 1.94, $CI_{95\%}$ [1.73, 2.15]. As expected, this target valence manipulation was considerably weaker than the manipulation used in Experiments 1–2 (recall that the effect sizes on the manipulation check were *d* = 2.97 and *d* = 3.11, respectively). Participants also indicated that Beth empathized with Ann to a greater extent when she gave an empathic (vs. non-empathic) response (*M* = 5.61, *SD* = 1.29 vs. *M* = 2.95, *SD* = 1.55), *t*(506) = 21.45, *p* < .001, *d* = 1.87, $CI_{95\%}$ [1.67, 2.08], and they rated Ann as feeling negative at the beginning of the interaction (*M* = 2.92, *SD* = 1.69), with the mean significantly below the mid-point of the scale, *t*(525) = 14.67, *p* < .001, *d* = 0.64, $CI_{95\%}$ [0.55, 0.73].

Experiment 3

Participants and Power Analysis

MTurk workers (N = 507, 52% female, 44% male, 4% no gender information; M_{age} = 37.4, SD_{age} = 12.6) participated online for modest remuneration. We determined our target sample size by running two power analyses based on the effect size estimates of the key interactions in Experiments 1 (around η_p^2 = .026) and 2 (around η_p^2 = .011). Detecting these two effect sizes at 80% power (α = .05) would require sample sizes of N = 296 and N = 708, respectively. Because we were unsure which effect size was more likely, and because we anticipated an exclusion rate of around 19% (based on the average exclusion rate in Experiments 1 and 2), we set a target sample size toward the higher end of the sample sizes suggested by the power analyses and collected data from N = 500. As in Experiments 1 and 2, we decided *a priori* on the following exclusion criteria: failing the attention check on Beth's response to Ann, failing the attention check on Ann's employer, and giving identical non-neutral responses (i.e., other than 4 on 7-point scales) across all dependent variables. The numbers of participants who met each criterion were n = 33, n = 65, and n = 6, respectively. The final sample was N = 416 (some participants met more than one criterion).

Manipulation Checks

All manipulations were successful: Participants evaluated Ann more positively when she was positively (vs. negatively) portrayed (M = 5.87, SD = 1.05 vs. M = 2.01, SD = 1.49), $t(323)$ = 29.87, $p < .001$, d = 3.05, $CI_{95\%}$ [2.76, 3.33]. They also indicated that Beth empathized with Ann to a greater extent when she gave an empathic (vs. non-empathic) response (M = 5.59, SD = 1.21 vs. M = 2.67, SD = 1.39), $t(404)$ = 22.80, $p < .001$, d = 2.24, $CI_{95\%}$ [1.99, 2.48]. In addition, participants rated Ann as feeling positive at the beginning of the interaction (M = 6.20, SD = 1.22), with the mean significantly above the mid-point of the scale, $t(415)$ = 36.67, $p < .001$, d = 1.80, $CI_{95\%}$ [1.64, 1.95].

Experiment 4

Participants and Power Analysis

We publicly pre-registered our analysis plan, including power analyses, target sample size, inclusion and exclusion criteria, and planned data analyses, on AsPredicted (http://aspredicted.org/blind.php?x=xu9ur5). MTurk workers (N = 838, 58% female, 42% male; M_{age} = 39.6, SD = 12.4) participated online for modest remuneration. We planned to power this experiment at 80% ($\alpha = .05$) to detect the expected effect size of response type on the primary dependent variables in the positive target condition. A conservative estimate of $\eta_p^2 = .031$ from simple effects analyses in the previous experiments suggested a target sample size of n = 124 per condition (N = 744). Based on an anticipated exclusion rate of 10% estimated from previous experiments, we set a target sample size of N = 820. We decided *a priori* to exclude participants based on the same three exclusion criteria from Experiment 3.[31] The numbers of participants who met each exclusion criterion were n = 45, n = 63, and n = 1, respectively. The final sample was N = 744 (some participants met more than one exclusion criterion).

Manipulation Checks

Following our pre-analysis plan, we conducted a one-tailed independent samples t-test on the manipulation check of target valence. Participants viewed Ann more positively when she was positively (vs. negatively) portrayed (M = 5.77, SD = 1.06 vs. M = 1.94, SD = 1.36), $t(680) = 42.53$, $p < .001$, d = 3.15, $CI_{95\%}$ [2.93, 3.36]. In addition, a one-tailed one sample t-test on Ann's affect confirmed that participants rated Ann as feeling negative at the beginning of the interaction (M = 2.35, SD = 1.37), with the mean significantly below the mid-point of the scale,

[31] We had reported an additional exclusion criterion in the pre-registration: excluding participants who fail the captcha verification at the beginning of the experiment. In reality, because the captcha verification appeared before any data could be recorded, all participants with recorded data passed the captcha verification.

$t(740) = 32.81$, $p < .001$, $d = 1.21$, $CI_{95\%}$ [1.11, 1.30]. Furthermore, an exploratory analysis indicated that, as in the pilot study, participants in the positive empathic response condition thought Beth empathized with Ann significantly more than did participants in both the positive non-empathic response condition ($M = 5.52$, $SD = 1.33$ vs. $M = 4.19$, $SD = 1.79$), $t(463) = 9.44$, $p < .001$, $d = 0.84$, $CI_{95\%}$ [0.66, 1.02], and the neutral non-empathic condition ($M = 2.82$, $SD = 1.43$), $t(480) = 21.51$, $p < .001$, $d = 1.95$, $CI_{95\%}$ [1.73, 2.16].

Fit of Latent Mediation Models

A summary of model fit indices is reported in Table S2.1.

Table S2.1

Summary of Latent Moderated Mediation Models Tested in Experiment 4.

Model	Planned or Exploratory?	Predictor	DV	χ^2	*CFI*	*TFI*	*RMSEA*
1	Planned	Empathy	Respect/liking	269.86	0.97	0.96	0.07
2	Planned		Warmth	358.81	0.95	0.94	0.08
3	Planned	Positivity	Respect/liking	513.89	0.94	0.92	0.10
4	Planned		Warmth	580.33	0.93	0.90	0.11
5	Exploratory	Empathic vs.	Respect/liking	272.33	0.95	0.94	0.09
6	Exploratory	positive non-empathic	Warmth	383.28	0.91	0.89	0.11

Note. Although the fit indices of some models slightly differed from conventional recommendations, inspection of residual matrices suggested that all models fit the data reasonably well and that the fit indices were oversensitive to minor model misspecifications, given the low unique variances of some observed variables (< .10; Browne, MacCallum, Kim, Andersen, & Glaser, 2002). In all models, $df = 57$, ps < .001.

Descriptions of Moderated Mediation Models 3–6

Moderated mediation models with response positivity as predictor (Models 3 and 4).

In Models 3 and 4, we conducted our planned moderated mediation analyses using response positivity as the predictor. Analysis on respect/liking (Model 3) indicated that response positivity significantly predicted the mediator, $a = 1.78$, $p < .001$, and that the mediator × target valence

interaction significantly predicted respect/liking, $b_{mod} = 0.54$, $p < .001$, suggesting the presence of second-stage moderation. However, because the response positivity × target valence interaction did not significantly predict the mediator, $a_{mod} = -0.08$, $p = .355$, there was no evidence of first-stage moderation. Supporting these results, the effect of response positivity on inferences about Beth's attitudes towards Ann was similar across target valence, $a_{pos} = 1.70$ vs. $a_{neg} = 1.86$, but the association between the mediator and respect/liking was stronger when Ann was positively portrayed, $b_{pos} = 0.97$ vs. $b_{neg} = -0.10$, and the overall indirect effect was also stronger when Ann was positively portrayed, $a_{pos}b_{pos} = 1.65$ vs. $a_{neg}b_{neg} = -0.18$.

We then conducted the same analysis on warmth (Model 4), and the results were similar. We again saw evidence of second-stage moderation, in which response positivity significantly predicted the mediator (*a*), and the mediator × target valence interaction significantly predicted warmth, $b_{mod} = 0.58$, $p < .001$. Because a_{mod} was not significant, there was again no evidence of first-stage moderation. Supporting these results, the effect of response positivity on inferences about Beth's attitudes toward Ann was similar across target valence, $a_{pos} = 1.68$ vs. $a_{neg} = 1.85$, but the association between the mediator and warmth was stronger when Ann was positively portrayed, $b_{pos} = 1.23$ vs. $b_{neg} = 0.07$, and the overall indirect effect was also stronger when Ann was positively portrayed, $a_{pos}b_{pos} = 2.06$ vs. $a_{neg}b_{neg} = 0.14$.

Taken together, Models 3 and 4 indicated that second-stage moderated mediation was present when we compared the effects of positive versus neutral responses: Inferences about Beth's attitudes toward Ann mediated the response positivity × target valence interaction on evaluations of Beth, but such inferences were predicted only by response positivity and did not differ by target valence.

Moderated mediation models with empathic vs. positive non-empathic response as predictor (Models 5 and 6). We explored within the empathic and positive non-empathic response conditions whether the response type × target valence interaction on evaluations of Beth was mediated. Analysis on respect/liking showed that the response type × target valence interaction significantly predicted the mediator, $a_{mod} = 0.25$, $p < .001$, and that the mediator significantly predicted respect/liking, $b = 0.37$, $p < .001$, suggesting the presence of first-stage moderation (Model 5). In addition, response type significantly predicted the mediator, $a = 0.18$, $p = .001$, and the mediator × target valence interaction significantly predicted respect/liking, $b_{mod} = 0.41$, $p < .001$, suggesting the presence of second-stage moderation as well. Supporting these results, the effect of response empathy on inferences about Beth's liking for Ann, the association between inferences about Beth's liking for Ann and respect/liking, and the overall indirect effect were all stronger when Ann was positively portrayed, $a_{pos} = 0.43$ vs. $a_{neg} = -0.06$, $b_{pos} = 0.78$ vs. $b_{neg} = -0.04$, $a_{pos}b_{pos} = 0.34$ vs. $a_{neg}b_{neg} = 0.00$.

The same exploratory analysis on warmth showed highly similar results: Response type × target valence interaction significantly predicted the mediator (a_{mod}), and the mediator significantly predicted warmth, $b = 0.57$, $p < .001$, suggesting the presence of first-stage moderation (Model 6). In addition, response type significantly predicted the mediator (a), and the mediator × target valence interaction significantly predicted warmth, $b_{mod} = 0.42$, $p < .001$, suggesting the presence of second-stage moderation as well. Supporting these results, the effects of the empathic (vs. positive non-empathic) response on inferences about Beth's attitudes toward Ann, the associations between the mediator and Beth's warmth, and the overall indirect effects were all stronger when Ann was positively portrayed, $a_{pos} = 0.43$ vs. $a_{neg} = -0.06$, $b_{pos} = 0.99$ vs. $b_{neg} = 0.14$, $a_{pos}b_{pos} = 0.43$ vs. $a_{neg}b_{neg} = -0.01$. Taken together, Models 5 and 6 indicated that

evidence of both first- and second-stage moderated mediation was present even when comparing only the effects of empathic versus positive non-empathic responses: Inferences about Beth's attitudes toward Ann mediated the response type × target valence interaction on evaluations of Beth, and such inferences were predicted by the response type × target valence interaction.

Experiment 5

Participants and Power Analysis

MTurk workers (N = 504, 50% female, 41% male, 9% no gender information; M_{age} = 39.0, SD_{age} = 11.9) participated online for modest remuneration. We powered our experiment to detect two effects: the effect of response type on respect/liking, and the indirect effect of the mediator on the dependent variables. Our experimental design was similar to that of the negative target conditions in Experiment 1; however, we reasoned that the condemning (vs. empathic) response should have a larger effect than the non-empathic (vs. empathic) response. Therefore, we estimated the effect size of response type on respect/liking as d = 0.28, which was twice as large as the size of the simple effect of non-empathic (vs. empathic) response on respect/liking in Experiment 1 (d = 0.14). Powering this experiment to detect an effect size of d = 0.28 at 80% (α = .05) requires N = 404. This sample size also affords >80% power to detect an indirect effect as small as $a \times b$ = 0.03, based on simulations using the power analysis app for mediation models developed by Schoemann, Boulton, and Short (2017).[32] Using a conservative estimate of 20% exclusion rate, we aimed for N = 500 participants. We decided *a priori* on the same three exclusion criteria used in Experiments 3 and 4. The numbers of participants who met each

[32] We conducted a power analysis using observed mediation models instead of our planned latent mediation models due to challenges of conducting power analysis for the latter. Because we anticipated that our latent variables would be highly reliable (αs = .90–.95), however, using latent variables in our mediation models should result in negligible power loss (see Table 3 in Ledgerwood & Shrout, 2011; also see Wang & Rhemtulla, 2021).

criterion were n = 22, n = 37, and n = 0, respectively. The final sample was N = 452 (some participants met more than one exclusion criterion).

Pilot Study

In order to ensure that our response type manipulation was successful, we conducted a pilot study on four candidate responses. In this pilot study (N = 201; 51% female, 39% male, 10% no gender information; M_{age} = 39.5, SD_{age} = 12.2), participants read the same instructions and Ann's experience as those in Experiment 4, but they did not learn any information about Ann. After reading what Ann said, participants then saw four responses from Beth presented in randomized order and rated how positive and how empathic each response was. The four responses were the three responses used in Experiment, as well as a neutral empathic response ("Okay, I can understand why you would feel stressed in this situation").

The positive empathic and positive non-empathic responses were comparably positive (M = 5.23, SD = 1.23 vs. M = 5.47, SD = 1.46), $t(200) = 1.94$, $p = .053$, $d = 0.14$, $CI_{95\%}$ [-0.06, 0.33], and more positive than the neutral non-empathic response (M = 3.14, SD = 1.19), ts > 17.85, ps < .001, ds > 1.25. The positive empathic response was also more empathic than both the positive non-empathic response (M = 5.89, SD = 1.23 vs. M = 3.90, SD = 1.91), $t(200) = 12.38$, $p < .001$, $d = 0.87$, $CI_{95\%}$ [0.67, 1.08], and the neutral non-empathic response (M = 2.46, SD = 1.38), $t(199) = 27.91$, $p < .001$, $d = 1.97$, $CI_{95\%}$ [1.73, 2.21]. Because the neutral empathic response was rated almost as positive (M = 4.74, SD = 1.08) as the two positive responses and significantly more positive than the neutral non-empathic response, $t(200) = 15.24$, $p < .001$, $d = 1.07$, $CI_{95\%}$ [0.87, 1.28], we did not use the neutral empathic in main experiment.

Manipulation Checks

All manipulations were successful: Participants viewed Ann negatively (M = 2.02, SD = 1.64), with the mean significantly below the scale mid-point, $t(451) = 25.59$, $p < .001$, d = 1.20, $CI_{95\%}$ [1.08, 1.32]. Participants in the empathic (vs. condemning) response conditions indicated that Beth empathized with Ann more (M = 5.62, SD = 1.26 vs. M = 1.51, SD = 1.01), $t(423) = 38.11$, $p < .001$, d = 3.60, $CI_{95\%}$ [3.30, 3.90]. Participants also rated Ann as feeling negative at the beginning of the interaction (M = 2.50, SD = 1.55), with the mean significantly below the scale mid-point, $t(451) = 20.53$, $p < .001$, d = 0.97, $CI_{95\%}$ [0.85, 1.08].

Experiment 6

Participants and Power Analysis

We publicly pre-registered our analysis plan, including power analyses, target sample size, inclusion and exclusion criteria, and planned data analyses, on AsPredicted (http://aspredicted.org/blind.php?x=89g5fh). MTurk workers (N = 566, 48% female, 52% male; M_{age} = 36.3, SD_{age} = 10.9) participated online for modest remuneration. We powered our experiment to detect a potential effect of response type × character gender on respect/liking. We estimated the main effect of response type on respect/liking to be d = 0.50 (a more conservative estimate than d = 0.72 as observed in Experiment 5), which required n = 64 per cell for 80% power (α = .05). The sample size per cell needed to detect a 2 × 2 between-subjects interaction that eliminates the main effect (a "knockout" interaction) is twice the sample size per cell needed to detect the main effect (Giner-Sorolla, 2018; Ledgerwood, 2019), suggesting a target sample size of N = 512. Based on an anticipated exclusion rate of 10% estimated from previous experiments, we aimed for N = 570 participants. We decided *a priori* on the same three exclusion criteria used in Experiments 3–5. The numbers of participants who met each criterion were n =

86, n = 115, and n = 3, respectively. The final sample was N = 404 (some participants met more than one exclusion criterion).

Manipulation Checks

All our manipulations were successful: Participants viewed the target negatively (M = 2.44, SD = 1.94), with the mean significantly below the scale mid-point, $t(403) = 16.25$, $p < .001$, d = 0.81, $CI_{95\%}$ [0.70, 0.92]. Participants in the empathic (vs. condemning) response conditions indicated that the responder empathized with the target more (M = 5.60, SD = 1.29 vs. M = 1.99, SD = 1.58), $t(387) = 25.17$, $p < .001$, d = 2.50, $CI_{95\%}$ [2.24, 2.77]. Participants also rated the target as feeling negative at the beginning of the interaction (M = 2.84, SD = 1.82), with the mean significantly below the scale mid-point, $t(403) = 12.85$, $p < .001$, d = 0.64, $CI_{95\%}$ [0.53, 0.75]. We explored whether character gender inadvertently affected any of the effects above; it did not, ps > .353.

Experiment 7

Participants and Power Analysis

We publicly pre-registered our analysis plan, including power analyses, target sample size, inclusion and exclusion criteria, and planned data analyses, on AsPredicted (http://aspredicted.org/blind.php?x=c8i59d). MTurk workers (N = 573, 52% female, 48% male; M_{age} = 36.9, SD_{age} = 11.8) participated online for modest remuneration. We planned to power this experiment at 80% ($\alpha = .05$) to detect the expected response type × disclosed experience interaction on respect/liking. Similar to Experiment 6, we estimated the main effect of response type on respect/liking to be d = 0.50 (a more conservative estimate than d = 0.72 as observed in Experiment 5), which required n = 64 per cell for 80% power ($\alpha = .05$). The sample size per cell needed to detect a 2 × 2 between-subjects interaction that eliminates the main effect (a

"knockout" interaction) is twice the sample size per cell needed to detect the main effect (Giner-Sorolla, 2018; Ledgerwood, 2019), suggesting a target sample size of $N = 512$. Based on an anticipated exclusion rate of 10% estimated from previous experiments, we aimed for $N = 570$ participants. We decided *a priori* on the same three exclusion criteria used in Experiments 3–5. The numbers of participants who met each exclusion criterion were $n = 42$, $n = 84$, and $n = 0$, respectively. The final sample was $N = 468$ (some participants met more than one criterion).

Manipulation Checks

All manipulations were successful: Participants viewed Ann negatively ($M = 2.08$, $SD = 1.61$), with the mean significantly below the mid-point of the scale, $t(467) = 25.77$, $p < .001$, $d = 1.19$, $CI_{95\%}$ [1.07, 1.31]. Participants in the empathic (vs. condemning) response conditions indicated that Beth empathized with Ann to a greater extent ($M = 5.63$, $SD = 1.28$ vs. $M = 1.53$, $SD = 1.18$), $t(461) = 35.91$, $p < .001$, $d = 3.32$, $CI_{95\%}$ [3.04, 3.60]. Participants also rated Ann as feeling negative at the beginning of the interaction ($M = 2.46$, $SD = 1.56$), with the mean significantly below the mid-point of the scale, $t(467) = 21.37$, $p < .001$, $d = 0.99$, $CI_{95\%}$ [0.88, 1.10]. Furthermore, an unplanned, exploratory analysis indicated that participants in the job stress (vs. cancer stress) condition thought Ann's experience was more attributable to the nature of her job ($M = 5.49$, $SD = 1.41$ vs. $M = 2.90$, $SD = 2.09$), $t(408) = 15.73$, $p < .001$, $d = 1.45$, $CI_{95\%}$ [1.25, 1.66]. In other words, we successfully manipulated how strongly the disclosed experience was linked to the source of target valence.

Confirmatory Factor Analysis Results in Experiments 3–7

In each of Experiments 3–7, we confirmed the two-factor structure of our dependent variables by conducting confirmatory factor analyses (CFA). Diagrams of the models are shown in Figure S2.2, and information on model fit is reported in Table S2.2. In each experiment, we

compared the two-factor model to a one-factor model in which all items loaded onto a single factor and found that the two-factor models provided superior fit (see Table S2.2).

Table S2.2

Summary of CFA Models Tested in Experiments 3–7.

Experiment	Model	χ^2	*CFI*	*TFI*	*RMSEA*	χ^2_{diff}
Experiment 3	Two-factor	103.31	0.98	0.97	0.10	
	One-factor	331.49	0.91	0.88	0.19	228.19
Experiment 4	Two-factor	151.26	0.98	0.97	0.10	
	One-factor	662.24	0.91	0.87	0.21	510.98
Experiment 5	Two-factor	115.70	0.98	0.96	0.11	
	One-factor	639.94	0.84	0.78	0.26	524.24
Experiment 6	Two-factor	115.87	0.97	0.96	0.11	
	One-factor	617.87	0.84	0.77	0.27	509.81
Experiment 7	Two-factor	163.78	0.97	0.95	0.13	
	One-factor	741.00	0.84	0.78	0.28	577.22

Note: In all two-factor models, $df = 19$, ps $< .001$; in all one-factor models, $df = 20$, ps $< .001$. For each experiment, χ^2_{diff} is the chi-square difference between the two-factor model and the one-factor model. In all chi-square difference tests, $df_{diff} = 1$, ps $< .001$.

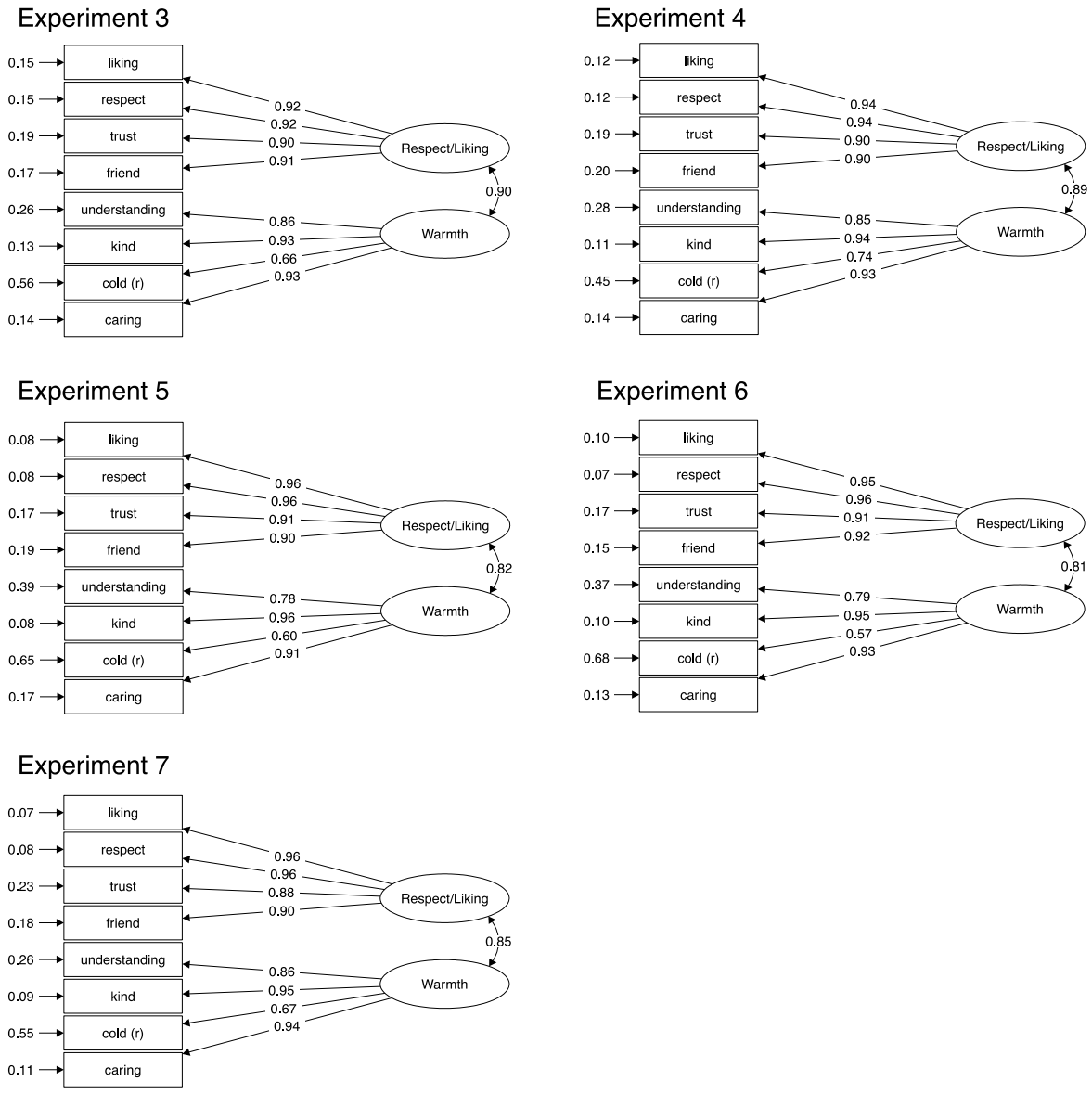

Figure S2.2. Diagram of the two-factor CFA models in Experiments 3–7. We set the variance of each latent variable to 1 in order to identify the scale of the model. All estimates are presented in standardized metric. The item "cold" was reverse-coded.

Full Statistical Models of the Latent Mediation Analyses in Experiments 4–7

We conducted latent moderated mediation analyses in Experiments 4 and 7 and latent simple mediation analyses in Experiments 5–7. Diagram of the full statistical model for latent moderated mediation analyses is presented in Figure S2.3, and diagram of the full statistical model for latent simple mediation analyses is presented in Figure S2.4.

In Figures S2.3 and S2.4, *inferences* are inferences about Beth's attitudes toward Ann. For respect/liking, items 1–4 indicate how much participants *liked*, *respected*, *trusted*, and *would like to be friends with* Beth; for warmth, items 1–4 indicate how *understanding*, *kind*, *cold* (reverse-coded), and *caring* Beth was. The items *like*, *positive*, and *unfavorable* indicate how much participants agreed that Beth liked Ann, felt positive toward Ann, and had an unfavorable opinion of Ann (reverse-coded). In both models, we allowed the residual covariance between the two positively-worded items of the mediator (agreement with the statements "[Responder] likes [target]" and "[Responder] feels positive toward [target]") to be freely estimated. We did so to reduce model misspecification of the mediator from ignoring wording-related covariance (Marsh, 1996) and better isolate the true mediator variance, which, in turn, should provide greater power and more accurate indirect effect estimates (Gonzalez & MacKinnon, 2020). To retain local independence of the latent mediator, we constrained the factor loadings of those two items to be equal. In the latent moderated mediation model, we additionally allowed the residual covariance between the two product indicators of the inferences × target valence latent variable that involve the two positively-worded items ("*like* × target valence" and "*positive* × target valence") to be freely estimated. We constrained the factor loadings of those product indicators to be equal as well. For visual simplicity, the residual variances of all variables and the covariances of all exogenous variables are omitted from the figures.

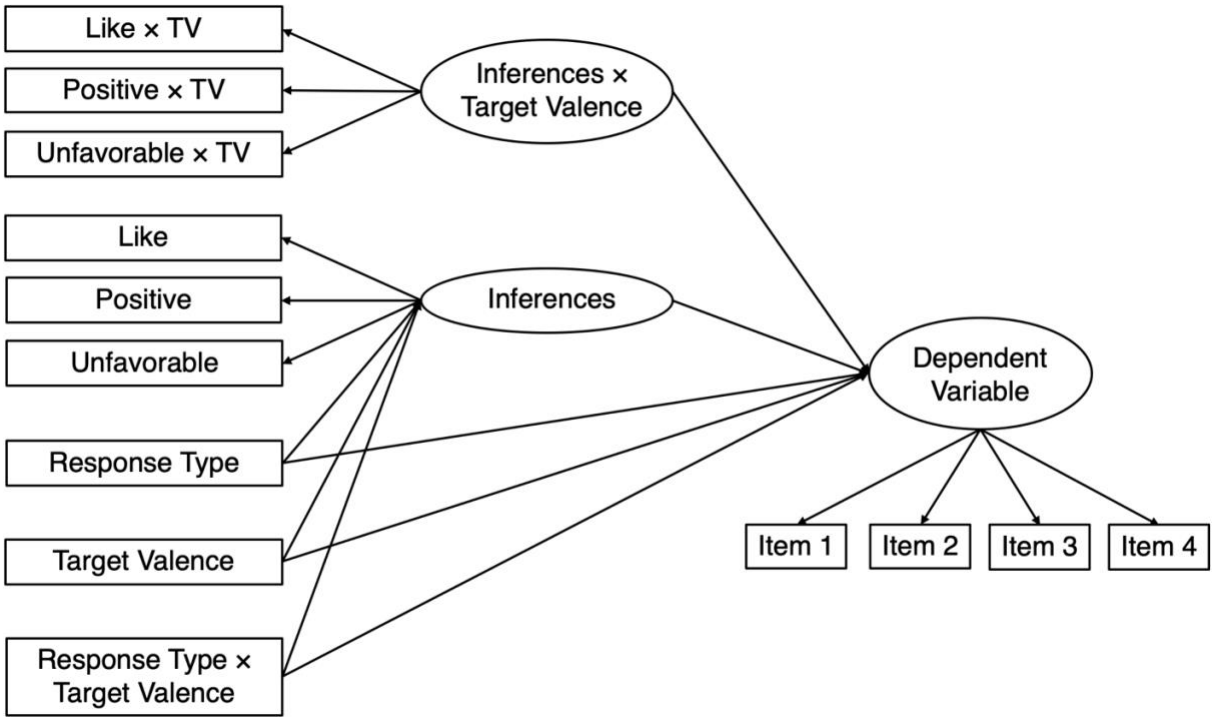

Figure S2.3. Full statistical model of the latent moderated mediation analyses in Experiments 4 and 7. TV = target valence. The latent interaction term inferences × target valence was measured by the product indicators that were created from the indicators of inferences and the observed target valence variable using the all-pairs approach (Foldnes & Hagvet, 2014; Wall & Amemiya, 2001). Details of the observed predictors (response type, target valence, and response type × target valence interaction) are reported in the paper.

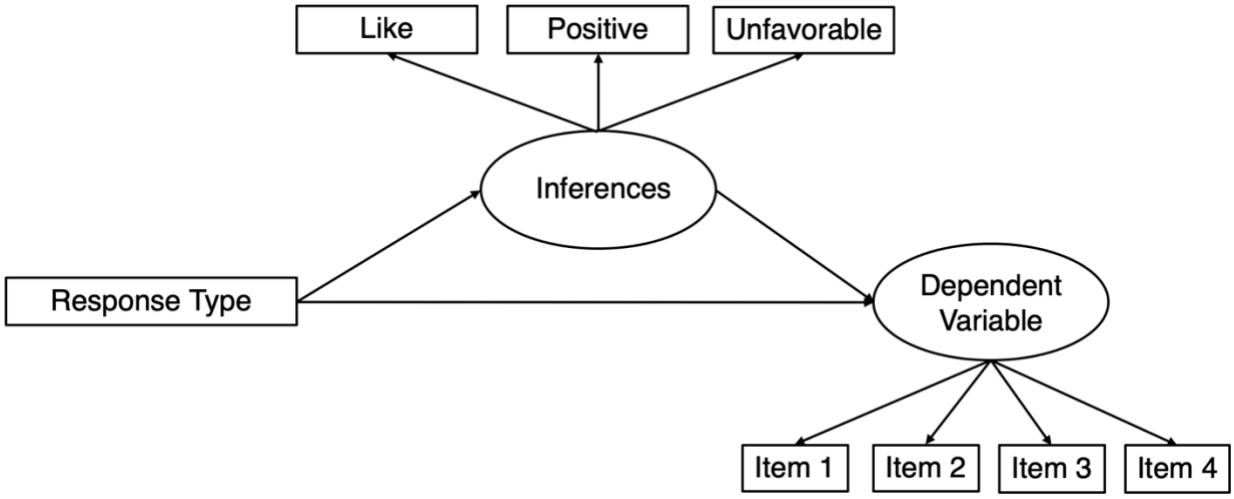

Figure S2.4. Full statistical model of the latent simple mediation analyses in Experiments 5–7. Details of response type are reported in the main text.

Mediational Evidence in Experiments 5–7

Fit of Simple Mediation Models

The fit indices of the simple mediation models using our main analytic approach are reported in Table S2.3.

Table S2.3

Fit Indices of the Latent Simple Mediation Models in Experiments 5–7 and the Pooled Data.

DV	Dataset	χ^2	*CFI*	*TFI*	*RMSEA*
Respect/liking	Experiment 5	164.62	0.97	0.95	0.13
	Experiment 6	122.60	0.97	0.95	0.12
	Experiment 7	97.29	0.96	0.94	0.14
	Pooled Data	335.56	0.97	0.95	0.13
Warmth	Experiment 5	160.66	0.96	0.93	0.13
	Experiment 6	233.08	0.92	0.87	0.17
	Experiment 7	86.18	0.95	0.93	0.13
	Pooled Data	412.45	0.95	0.92	0.14

Note. Although the fit indices of some models slightly differed from conventional recommendations, inspection of residual matrices suggested that all models fit the data reasonably well and that the fit indices were oversensitive to minor model misspecifications, given the low unique variances of some observed variables (< .10; Browne et al., 2002). In all models, $df = 18$, ps < .001.

Alternative Analytic Approaches to Simple Mediation Models in Experiments 5–7

To assess the robustness of evidence for indirect effects from the simple mediation analyses, we compared the results with those from two alternative analytic approaches. The first analytic approach was almost identical to the main approach but ignored wording differences among the items in the mediator (i.e., the mediator items had freely estimated factor loadings and independent residual variances). This approach reflects our originally intended analytic strategy but resulted in worse model fit across all datasets ($\Delta\chi^2$s = 9.78–42.51). The indirect effect estimates from this first alternative approach had the same level of significance and signs as the estimates from the main analytic approach in Experiments 5 and 7, but not for Experiment 6. The indirect effect from the pooled data was not significant for respect/liking, $ab = 0.07$, $p = .402$, $CI_{95\%}$ [-0.10, 0.24], or warmth, $ab = -0.15$, $p = .090$, $CI_{95\%}$ [-0.32, 0.02], with the indirect effect

estimate for warmth marginally significant but in the opposite direction as that estimated from the main analytic approach.

The second alternative approach contained only observed (rather than latent) variables and modeled both the mediator and the DVs as composite scores. The indirect effect estimates from this second alternative approach had the same level of significance and signs as the estimates from the main analytic approach for all datasets, including the estimates from the pooled data for both respect/liking, $ab = 0.50$, $CI_{95\%}$ [0.24, 0.77], and warmth, $ab = 0.08$, $CI_{95\%}$ [-0.16, 0.32].

Moderated Mediation Models in Experiment 7

We conducted moderated mediation analyses on data from Experiment 7 by entering disclosed experience as a moderator (+1/2 = job stress, -1/2 = cancer stress). The mediation model for respect/liking had acceptable fit, $\chi^2(57) = 772.11$, $p < .001$, $CFI = 0.89$, $TLI = 0.85$, $RMSEA = 0.16$.[33] The response type × disclosed experience interaction predicted the mediator, $a_{mod} = 0.63$, $p = .007$, but the mediator only marginally predicted respect/liking, $b = 0.11$, $p = .064$, suggesting no evidence of first-stage moderation, $a_{mod}b = 0.07$, $p = .126$. There was no evidence of second-stage moderation either, $ab_{mod} = 0.11$, $p = .526$. These results suggest that the effect of response type on inferences about Beth's attitudes toward Ann was stronger when Ann disclosed cancer (vs. job) stress, $a_{cancer} = -4.10$ vs. $a_{job} = -3.47$, but the associations between inferences about Beth's attitudes toward Ann and respect/liking were comparable across disclosed experience, $b_{cancer} = 0.12$ vs. $b_{job} = 0.09$. The overall indirect effect was also

[33] We concluded that the latent moderated mediation models in Experiment 7 provided acceptable fit after a holistic assessment of the fit indices as well as the residual matrices and modification indices of these models. Although the fit indices here are less than ideal, we observed low unique variances (< .10) similar to ones observed in Experiment 4 for the majority of the observed variables in these models. These low unique variances suggest that most items were highly reliable (e.g., items on how much participants liked and respected Beth both had standardized factor loadings above .96) and might have led to fit indices that were oversensitive to minor model misfit (Browne et al., 2002). Furthermore, we did not identify any conceptually sensible modification to these models that would non-trivially improve model fit, and none of the key parameter estimates changed in significance when we explored conceptually sensible modifications. Thus, we interpret these models as they were specified in our pre-analysis plan.

comparable across disclosed experience, $a_{cancer}b_{cancer}$ = -0.51 vs. $a_{job}b_{job}$ = -0.32 (see Table S2.4 for all parameter estimates).

Table S2.4

Parameter Estimates and 95% Confidence Intervals from the Latent Moderated Mediation Models in Experiment 7.

Parameter	Respect/liking	Warmth
a	-3.78 [-4.27, -3.29]	-3.78 [-4.27, -3.29]
a_{job}	-3.47 [-3.98, -2.96]	-3.47 [-3.98, -2.96]
a_{cancer}	-4.10 [-4.67, -3.53]	-4.10 [-4.67, -3.53]
a_{mod}	0.63 [0.18, 1.08]	0.63 [0.18, 1.08]
b	0.11 [-0.01, 0.22]	0.21 [0.09, 0.33]
b_{job}	0.09 [-0.03, 0.22]	0.17 [0.04, 0.30]
b_{cancer}	0.12 [-0.00, 0.25]	0.24 [0.11, 0.37]
b_{mod}	-0.03 [-0.12, 0.06]	-0.07 [-0.17, 0.03]
c	-0.19 [-0.37, 0.00]	-0.94 [-1.14, -0.73]
c'	0.23 [-0.25, 0.70]	-0.16 [-0.65, 0.33]
$a_{job}b_{job}$	-0.32 [-0.76, 0.11]	-0.59 [-1.05, -0.14]
$a_{cancer}b_{cancer}$	-0.51 [-1.02, 0.01]	-0.99 [-1.54, -0.44]

The mediation model for warmth also had acceptable fit, $\chi^2(57) = 806.03$, $p < .001$, *CFI* = 0.87, *TLI* = 0.83, *RMSEA* = 0.17. In addition to the effect of the response type × disclosed experience interaction on the mediator, a_{mod} = 0.63, p = .006, the mediator predicted warmth, b = 0.21, p = .001, suggesting first-stage moderation, $a_{mod}b$ = 0.13, p = .034. There was no evidence of second-stage moderation, ab_{mod} = 0.26, p = .154. These results suggest that, although the associations between inferences about Beth's attitudes toward Ann and warmth were comparable across disclosed experience, b_{cancer} = 0.24 vs. b_{job} = 0.17, the overall indirect effect was stronger when Ann disclosed cancer (vs. job) stress, $a_{cancer}b_{cancer}$ = -0.99 vs. $a_{job}b_{job}$ = -0.59 (see Table S2.4 for all parameter estimates).

Exploratory Analyses on Perceived Similarity Between Responder and Target

In all experiments, we included a single-item exploratory measure on perceived similarity between the responder and the target ("To what extent do you think [responder's name] and [target's name] are similar to each other?" 1 = *not at all*, 7 = *very much*). We conducted a series of analyses to explore the possibility that perceived similarity underlies evaluations of empathizers—namely, that participants in our experiments evaluated empathizers based on how similar they think the empathizer and the target is.

First, we explored whether the response type × target valence interaction effects that we observed for evaluations of empathizers in Experiments 1–4 are present for perceived similarity. Across the experiments, the response type × target valence ANOVAs showed main effects of response type but no response type × target valence interaction (except for a small interaction in Experiment 4; see Table S2.5 for results from each experiment). Given that the response type × target valence interaction on similarity was largely absent in Experiments 1–4, we conclude that similarity is unlikely to have driven evaluations of empathizers in those experiments.

Table S2.5

ANOVA on perceived similarity in Experiments 1–4.

Experiment	Predictor	df_n	df_d	F	p	η_p^2	$CI_{90\%}$
Experiment 1	Response Type	1	322	66.16	< .001	.17	[.11, .23]
	Target Valence	1	322	12.64	< .001	.04	[.01, .08]
	Interaction	1	322	0.47	.493	< .01	[.00, .02]
Experiment 2	Response Type	1	522	248.34	< .001	.32	[.27, .37]
	Target Valence	1	522	10.88	.001	.02	[.01, .04]
	Interaction	1	522	0.05	.831	< .01	[.00, .00]
Experiment 3	Response Type	1	412	340.78	< .001	.45	[.40, .50]
	Target Valence	1	412	3.28	.071	.01	[.00, .03]
	Interaction	1	412	1.45	.229	< .01	[.00, .02]
Experiment 4	Response Type	2	734	70.69	< .001	.16	[.12, .20]
	Target Valence	1	734	0.03	.853	< .01	[.00, .00]
	Interaction	2	734	5.12	.006	.01	[.00, .03]

Note. df_n = degree of freedom numerator; df_d = degree of freedom denominator; interaction = response type × target valence interaction.

Next, we explored the indirect effects of similarity. We compared two analytic approaches: With the first approach, we combined similarity with the three items on inferences about the responder's attitudes toward the target to form a four-item latent variable ("affinity"). We then tested for the indirect effects of this latent variable. With the second approach, we tested for the effects of similarity as a single-item mediator in mediation models where all variables were observed. Our mediation analyses focused on Experiments 4–7 to compare results from these two approaches with those from our planned analyses (and because the ANOVA results for similarity showed clear divergence from the results for our primary DVs in Experiments 1–3).

Estimates of the indirect effects are reported in Figure S2.5 (Experiment 4) and S2.6 (Experiments 5–7). The first approach yielded results that are highly similar to those from our planned approach (for which the mediator is inferences about the responder's attitudes toward the target): Almost all estimates from the first approach have the same signs, levels of significance, and largely overlapping 95% CIs as those from our planned approach. These findings suggest that similarity added little unique variance that accounted for the relations between the predictors and the DVs. Results from the second approach showed a more mixed picture. In almost all models with respect/liking as the DV and some models with warmth as the DV, estimates from the second approach largely match estimates from the other two approaches in terms of magnitude and signs. In the other models, estimates from the second approach diverge from those from our planned approach, but the direction of divergence differed across models (e.g., the second-stage moderated mediation effect from Model 4 in Experiment 4 was smaller but in the same direction; the mediation effect on warmth in Experiment 7 was in the opposite direction). Because these differences only appear in some models and experiments and

do not seem consistent, and because using observed versus latent variables tend to yield estimates that are more precise but less accurate (Ledgerwood & Shrout, 2011; Wang & Rhemtulla, 2021), we hesitate to draw substantive conclusions from these differences.

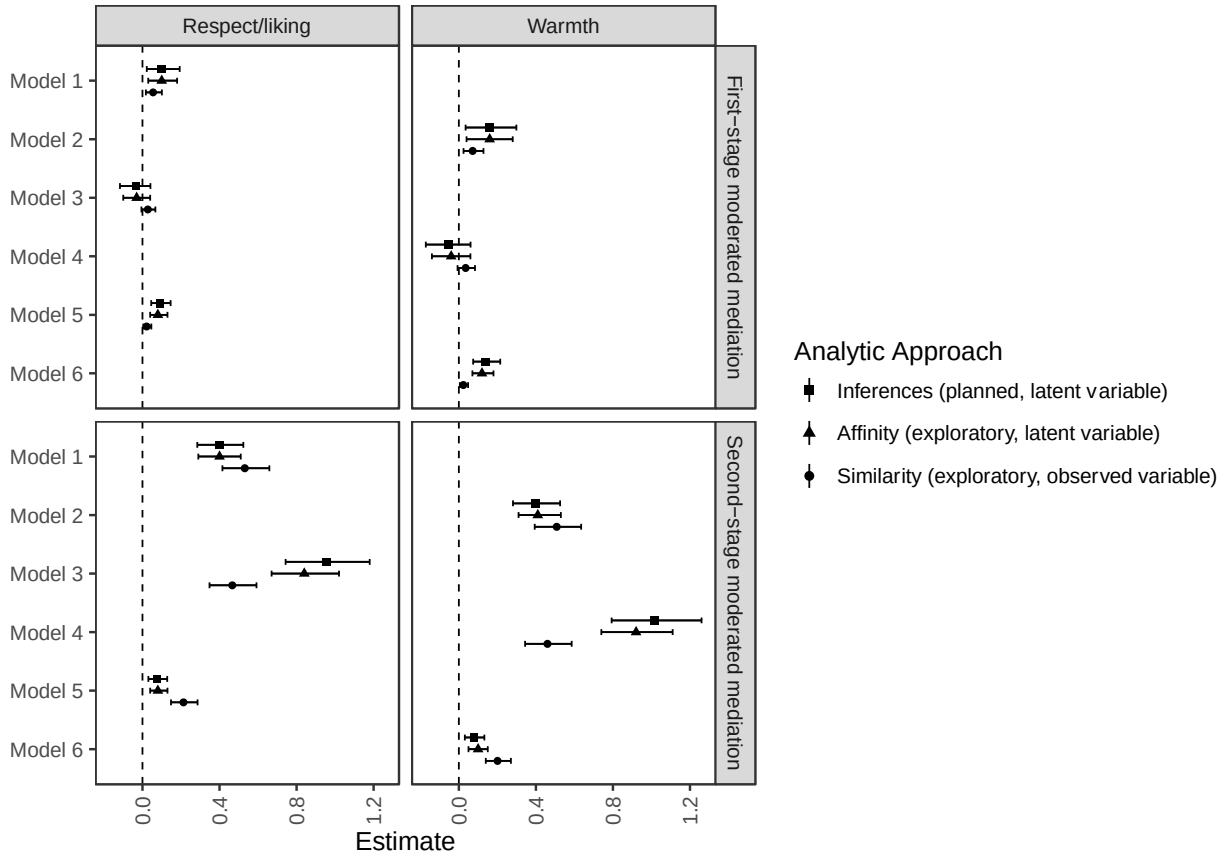

Figure S2.5. Indirect effect estimates from moderated mediation analyses in Experiment 4. Respect/liking was the DV in Models 1, 3, and 5; warmth was the DV in Models 2, 4, and 6 (see main text for details on these models). First-stage moderated mediation effect was $a_{mod}b$; second-stage moderated mediation effect was ab_{mod}. Results from the planned approach are also reported in the main text; we reproduce them here for ease of comparison.

Taken together, we speculate that our similarity item functioned like the inferences measure, in that they both seem to capture participants' thoughts about the relationship between the responder and the target (i.e., their affinity). We also note that the meaning of our similarity measure might have been ambiguous: Participants could have interpreted the item as asking whether the responder and the target had similar backgrounds (e.g., both working for a children's hospital or a white supremacist organization), or more broadly as whether they share certain

values or even demographic characteristics. We believe future research could more fruitfully and rigorously test for similarity as a mechanism by using better measures and manipulating both perceived similarity and inference affordance.

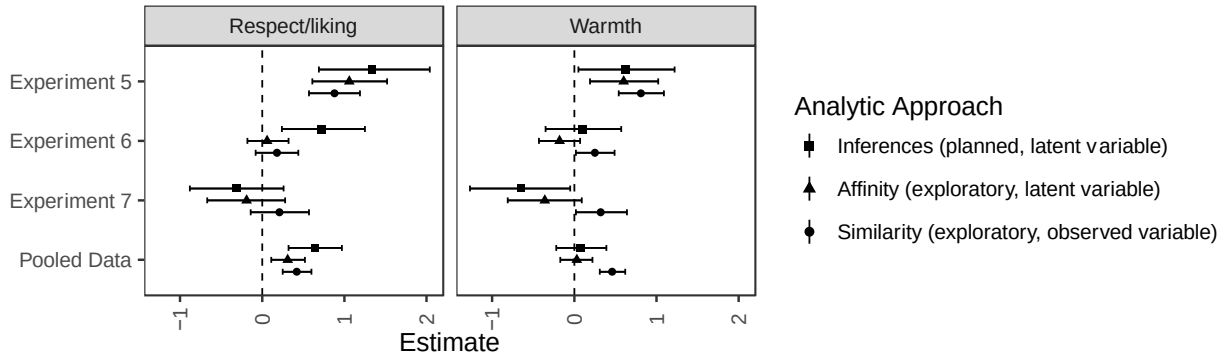

Figure S2.6. Indirect effect (*ab*) estimates from simple mediation analyses in Experiments 5–7 and the pooled data (see main text for details on these datasets). Results from the planned approach are also reported in the main text; we reproduce them here for ease of comparison.

Chapter 3

Power Analysis for Parameter Estimation in Structural Equation Modeling: A Discussion and Tutorial

Cite: Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*, 1–17.

Abstract

Despite the widespread and rising popularity of structural equation modeling (SEM) in psychology, there is still much confusion surrounding how to choose an appropriate sample size for SEM. Currently available guidance primarily consists of sample size rules of thumb that are not backed up by research, and power analyses for detecting model misfit. Missing from most current practices is power analysis to detect a target effect (e.g., a regression coefficient between latent variables). In this paper we (a) distinguish power to detect model misspecification from power to detect a target effect, (b) report the results of a simulation study on power to detect a target regression coefficient in a 3-predictor latent regression model, and (c) introduce a Shiny app, pwrSEM, for user-friendly power analysis for detecting target effects in structural equation models.

Keywords: power analysis, structural equation modeling, simulation, latent variables

Introduction

Structural equation modeling (SEM) is increasingly popular as a tool to model multivariate relations and to test psychological theories (e.g., Bagozzi, 1980; Hershberger, 2003; MacCallum & Austin, 2000). Despite this popularity, there remains a great deal of confusion about how to design an SEM study to be adequately powered. Empirical articles reporting SEM analyses rarely describe how sample size was determined, so it is unclear to what extent sample size planning is considered (Jackson et al., 2009). When researchers do report rationale for sample size, they often rely on rules of thumb that recommend either absolute minimum sample sizes (e.g., N = 100 or 200; Boomsma, 1982, 1985) or sample sizes based on model complexity (e.g., n = 5 to 10 per estimated parameter, Bentler & Chou, 1987; n = 3 to 6 per variable, Cattell, 1978). However, these rules of thumb do not always agree with each other, have little empirical support (Jackson, 2003; MacCallum & Austin, 2000; MacCallum et al., 1999), and generalize to only a small range of model types (Marsh et al., 1998).

The difficulty of forming sample size recommendations reflects in part the flexibility of SEM that eludes a one-size-fits-all solution, but it also reflects ambiguity about the goals associated with sample size planning: Researchers often want to have enough power in their studies to know both whether their model describes the data well, and whether specific effects within their model exist. In addition, and unrelated to power, researchers need a sample size that is large enough to ensure a stable model (i.e., to ensure that the estimation algorithm will converge on a solution). Yet the minimum sample size for model convergence is different than the minimum required to detect model misspecification, which will yet be different than the minimum required to detect a target effect within the model (Wolf et al., 2013).[34] A single

[34] Although not the focus of this paper, it is worth noting that another goal of sample size planning in SEM might be to obtain accurate estimates (e.g., Wolf et al., 2013).

analysis cannot reveal the minimum sample size that will achieve all of these disparate goals. We next explain the two types of power that are relevant to SEM.

Power to Detect a Misspecified Model versus Power to Detect a Target Effect

Two distinct modeling goals in SEM entail two different kinds of power. One goal is to determine how well the model as a whole describes the data; this goal requires that an analysis has enough power to detect a meaningful level of model misspecification. Another goal is to determine whether specific effects in the model exist—for example, whether one latent variable predicts another; this goal requires that an analysis has enough power to detect a minimally interesting effect size corresponding to a particular model parameter (Hancock & French, 2013; Lai & Kelley, 2011; Kaplan, 1995). These two types of power may require very different sample sizes, such that an SEM analysis may be well-powered to detect a model misspecification but poorly powered to detect a key effect, or vice versa.

Power to detect misspecification. Power to detect a misspecified model is the probability of correctly rejecting an incorrect model, given a specific degree of population misfit. Any structural equation model with positive degrees of freedom entails a set of hypotheses about the relations among variables that put constraints on the population covariance matrix. Fitting the model to data allows researchers to test the hypothesis that the covariance matrix implied by the model is equivalent to the covariance matrix in the population. The effect size to be detected is the degree of true model misfit, which summarizes the degree of discrepancy between the model-implied and population covariance matrices. As such, the model misfit effect size does not refer to any particular effect of interest *within* the model; rather, it is a global metric of how well (or rather, how poorly) a model describes data.

Many methods have been developed over the years for power analysis in this context, such as Satorra and Saris' (1985) chi-square likelihood ratio test comparison of exact fit of an incorrect null model and a correct alternative model (also see Mooijaart, 2003; Yuan & Hayashi, 2003), MacCallum, Browne, and Sugawara's (1996) root mean square error of approximation (*RMSEA*) tests of close and not-close fit, and extensions of this methods to other fit indices (e.g., Kim, 2005; MacCallum, Browne, & Cai, 2006; MacCallum & Hong, 1997). Several tutorials and online calculators for this type of power analysis are available (e.g., Hancock & Freeman, 2001; Preacher & Coffman, 2006; Zhang & Yuan, 2018).

Power to detect a target effect. Power to detect a target effect is the probability of correctly rejecting the null hypothesis that a key effect is zero in the population, given a specific true effect size. For example, a researcher might want to know whether latent variable X predicts latent variable Y. In this case, the effect size is the true parameter value (e.g., the size of the regression coefficient). Researchers might find this type of power more familiar and intuitive, because it parallels power analyses for *t*-tests, ANOVAs, and multiple regressions, and effect size metrics are comparable (e.g., regression coefficients that can be standardized and converted to common metrics like *r* and Cohen's *d*). Given a true effect of a particular size in the population, power to detect the effect is the probability that the estimated regression coefficient is significantly different from 0.[35] A "target effect" in this context can extend beyond structural regression coefficients to any model parameter that can be estimated, such as factor loadings,

[35] There are two possible tests of this null hypothesis. One is to compute the Wald test statistic by dividing the parameter estimate by its standard error and comparing the result to a z-distribution to get a *p*-value. The other is to fit a second model in which the parameter is constrained to zero and compare the two model test statistics using a likelihood-ratio test (i.e., a Chi-square difference test) with 1 degree of freedom. The latter method is preferred because it is invariant to differences in model identification (i.e., it gives exactly the same result no matter how the latent variable scaling is achieved; Griffin & Gonzalez, 2001). However, the differences are typically small, and for the purpose of power analysis, the Wald test method is computationally simpler and allows for the simultaneous testing of multiple parameters. The results in this paper, and the Shiny app, rely on the Wald test.

means, or residual covariances. Power analysis to detect a target effect in SEM can be challenging: Researchers must specify not only the value of the target parameter (i.e., its effect size), but also the values of all parameters in the population model. Existing guides require programming knowledge and are either limited to proprietary software (e.g., Mplus; Muthén & Muthén, 2002) or scattered across online resources, creating barriers for researchers to use them.

We seek to lower those barriers in this paper. First, we explain how power to detect a target effect is influenced by features of the model, and how power in SEM differs from power in multiple regression. We demonstrate these effects with a simulation study. Next, we discuss how to conduct a power analysis to detect a target effect, and we introduce pwrSEM, a point-and-click Shiny app that allows users to run power analysis to detect a target effect in SEM. We walk readers through the app in a hands-on guide with an example. Throughout the tutorial, we provide practical guidance regarding the choices involved in conducting a power analysis.

Factors Affecting Power to Detect a Target Effect

The factors that affect power to detect a target effect in SEM include well-known factors that affect power in any method, like sample size and effect size, as well as less familiar considerations like number of indicators, indicator reliability, and the values of other parameters in the model.[36] In this section, we briefly review how these factors affect power.

First and most obviously, larger effect sizes lead to greater power. The Wald test statistic of the null hypothesis that any SEM parameter is zero is the value of the parameter estimate divided by its standard error. As effect size (i.e., the absolute value of the true target parameter) increases, the average estimated value of the parameter increases, resulting in greater test statistics on average.

[36] These less familiar factors also affect power in ANOVA/regression (see Maxwell, 2000, for a discussion).

Second, having more information leads to greater power because it increases the *efficiency* of parameter estimation. Efficiency relates to the variability of the parameter estimate across repeated samples: Smaller sampling variability means that a parameter is more precisely estimated, so it will have (on average) a smaller estimated standard error (SE) and confidence interval, and thus a test to detect its difference from zero will be more powerful. The most straightforward way to increase information is to have a larger sample size. Other factors that affect efficiency include completeness of data (more missing values means less information, resulting in larger SEs and lower power) and distribution of data (e.g., data that are not multivariate normally distributed, such as ordinal variables and variables with high multivariate kurtosis, also lead to less efficient estimation; Savalei, 2014).

Third, power to detect parameters of a structural model (e.g., latent regression coefficients) is influenced by features of the measurement model, particularly the number and reliability of indicators. In the simple case of a model with a single latent predictor and a single latent outcome, power increases as a function of the *coefficient of determination* (also known as *maximal reliability* or *coefficient H*), which can be understood as the proportion of variation in the set of indicators that is explained by the latent variable. Alternatively, it can be viewed as the reliability of an optimally-weighted sum score computed from the items (Bollen, 1989; Dolan, Wicherts, & Molenaar, 2005; Hancock & French, 2013; Penev & Raykov, 2006). The coefficient of determination can be increased by adding more indicators to the measurement model and/or by increasing the reliability of the existing indicators.

Fourth, power to detect non-zero parameters in SEM may be affected by the size of the structural model (i.e., the number of latent variables that are modeled), the values of all the other structural paths, and the number of estimated paths. Although some suggestive evidence exists

(e.g., power varies as a function of the number of latent variables in confirmatory factor analyses; Wolf et al., 2013), we do not know of any studies that have attempted to systematically examine how estimation efficiency of one parameter is affected by these factors.

Clearly, a larger number of factors can influence power to detect a target effect in SEM. Whereas the effects of some of these factors are well understood, there is no easy way to identify a sample size requirement that guarantees sufficient power for any given parameter in any given model. In the next section, we use a simulation study to illustrate how some of these factors come together to affect power for a specific target parameter in a specific model.

Simulations: Power to Detect a Target Effect

Using simulated data from a simple SEM in which a latent variable *Y* is regressed on 3 latent predictors, *X*, *W*, and *Z*, we show how power to detect the regression coefficient of *Y* on *X* varies as a function of sample size and characteristics of the model.

Disclosures

Data and code for this paper are available via the Open Science Framework at https://osf.io/h8yfk/. The OSF project at this URL contains R code to reproduce the simulation study, results of the simulation study, and source code for the Shiny web app pwrSEM. The simulation study was conducted in RStudio (Version 1.2.1335; RStudio Team, 2018) and R (Version 3.6.0; R Core Team, 2019) with the R package lavaan (Version 0.6-5; Rosseel, 2012). pwrSEM was created in RStudio and R with the R packages lavaan, rhandsontable (Version 0.3.7; Owen, 2018), shiny (Version 1.3.2; Chang, Cheng, Allaire, Xie, & McPherson, 2019), semPlot (Version 1.1.1; Epskamp, 2015), semTools (Version 0.5-1; Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2018), and tidyr (Version 0.8.3; Wickham & Henry, 2019). All

simulations conducted are reported. Additional supplemental material will be available on the journal's website.

Method

We generated data from the model depicted in Figure 3.1, varying the values of the inter-correlations among latent predictors, factor loadings, number of indicators, and latent regression coefficients. For each condition, we simulated 10,000 datasets each for sample sizes ranging from 50 to 1000 and generated a power curve. Full details of the simulation design and method are reported in the supplemental material.

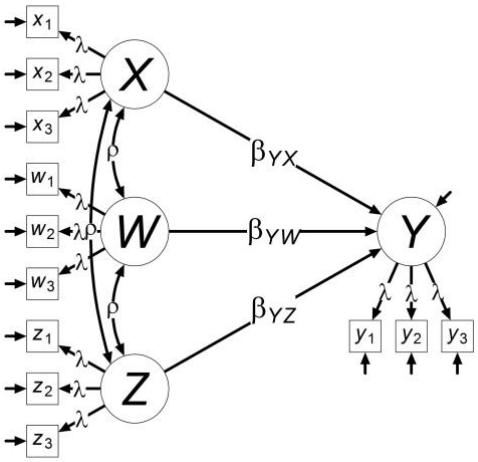

Figure 3.1. Population model to generate simulated data ($p/f = 3$ shown only). Variables shown in circles are latent; variables shown in squares are observed indicators. The target parameter for which power was estimated is β_{YX}. Labeled paths are parameters that were varied in the simulation. Indicator residual variances were set to $1 - \lambda^2$ such that all observed variables had unit variance. Residual variance of Y was also fixed such that the total variance of $Y = 1$.

We fit two models to each generated dataset (see Figure 3.2). The first was a structural equation model that corresponded to the population generating model. The second was a multiple regression model based on composite scores formed by summing the indicators of each latent factor. For each fitted model, we recorded (a) whether the estimation algorithm successfully

converged on a proper set of parameter estimates and was able to estimate standard errors, and (b) whether the *p*-value associated with $H_0: \beta_{YX} = 0$ was less than .05.

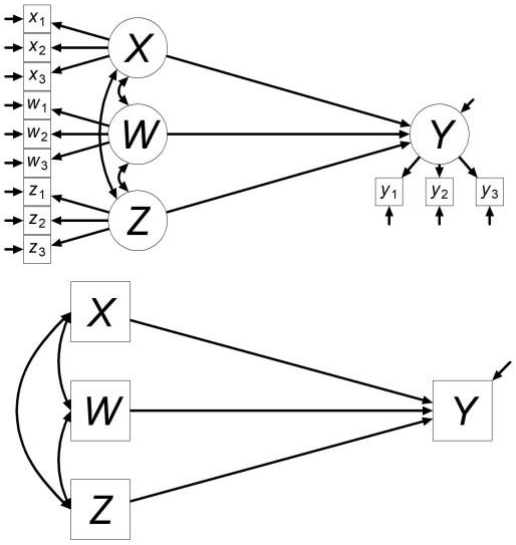

Figure 3.2. Fitted models. Top: fitted structural equation model (*p/f* = 3 shown only). Bottom: multiple regression model fitted to composite scores computed as sums of the set of indicators of each latent variable. All paths correspond to freely estimated parameters, except for latent variable (residual) variances, which were fixed to 1 for model identification.

Results

First, we examined model convergence rates and found predictable results: Serious convergence problems arose when the number of indicators was small, item reliabilities were low, and sample size was small. Full results are available in the supplemental material.

Next, we computed power as the proportion of converged cases in each condition in which the estimated regression coefficient of $X \rightarrow Y$ was significantly different from 0 ($\alpha = .05$).[37] Figure 3.3 displays the results for a single fixed set of values of the non-target structural

[37] We also checked whether the pattern of results was different if we calculated power as the proportion of all 10,000 simulations per condition in which the estimated regression coefficient of *X* on *Y* was significantly different from 0 ($\alpha = .05$), where nonconverged models counted as not rejecting H_0. Because non-convergence only substantially affected conditions in which the number of samples with β_{YX} significantly different from 0 was low, the pattern of results was largely unaffected (see Figure S3.2 in the supplemental material), with power in those conditions estimated as lower than presented here.

path coefficients: The latent predictor correlations (ρ) were .3, and the regression coefficients were β_{YW} = .1 and β_{YZ} = .2.

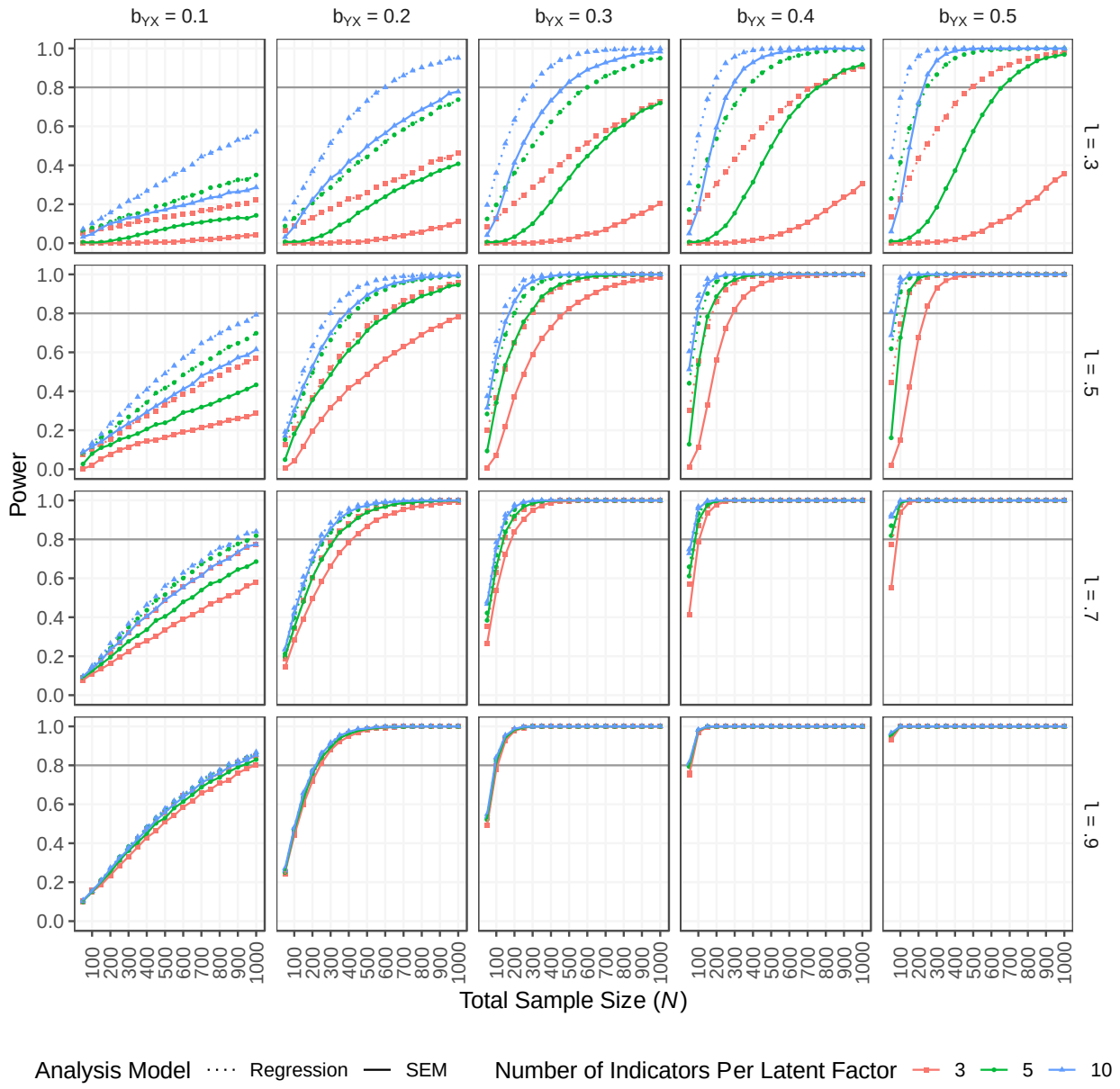

Figure 3.3. Power (y-axis) as a function of the total sample size (*N*; x-axis), population effect size (β_{YX}; horizontal panels), factor loading strength (λ; vertical panels), number of indicators per latent factor (*p*/*f*; line color and point shape), and analysis model (SEM or composite score regression; line type).

As expected, power increased with increasing sample size, effect size, item reliability (factor loadings), and scale length (number of indicators). To understand the effects of the

measurement model parameters (i.e., number of indicators and factor loadings) on power, recall that these two factors together determine the coefficient of determination (CD), which in turn affects power. Table 3.1 gives the coefficient of determination corresponding to each combination of measurement model parameters. Because the coefficient of determination can be interpreted as the reliability of an optimally-weighted sum score of the items, it can be approximated by the reliability of the (unweighted) scale when the items have similar reliability.

Table 3.1

Coefficient of Determination (CD) Corresponding to Each Combination of Measurement Model Parameters

Factor loading	Number of indicators		
	3	5	10
0.3	0.23	0.33	0.50
0.5	0.50	0.63	0.77
0.7	0.74	0.83	0.91
0.9	0.93	0.96	0.98

When the effect size was small ($\beta_{YX} = .1$, leftmost column of Figure 3.3), the only conditions to achieve .80 power in the latent variable SEM analysis were those with 5 or more highly reliable ($\lambda \geq .9$) items per factor (i.e., when CD $\geq .96$) and very large samples ($N \geq 900$). At more reasonable levels of scale reliability, power was quite low. For example, with 5 indicators loading at $\lambda = .7$(CD = .83), power just reached .50 at $N = 650$.

When the effect size was moderate ($\beta_{YX} = .3$, middle column of Figure 3.3), .80 power was typically attained at reasonably small sample sizes (as small as $N = 200$ when the CD reached .74 (e.g., when the measurement model contained 3 indicators loading at $\lambda = .7$ or 10 indicators loading at $\lambda = .5$). With moderate item reliabilities and 5 items per factor (CD = .63), $N \geq 300$ was required to attain power of .80. Notably, 3 or 5 unreliable indicators ($\lambda = .3$, CD $\leq .33$) per factor did not produce sufficient power to detect a medium effect size even at our

largest sample size (N = 1000). When the effect size was large ($\beta_{YX} = .5$, rightmost column of Figure 3.3), .80 power was attained whenever CD $\geq .5$ (i.e., with 3 or more indicators when $\lambda \geq .5$ or with 10 indicators when $\lambda = .3$) and $N \geq 250$.

These findings exemplify the ways in which effect size, sample size, and reliability of the measurement model contribute to power to detect a target effect. Additional figures in the supplemental material compare these results to results for different values of the structural parameters, including when the correlations among latent factors were .5 instead of .3 (Figure S3.3), and when the non-target regression coefficients were both .3 instead of .1 and .2 (Figure S3.4). As these supplemental figures demonstrate, it is very difficult to extract general principles about the effect of non-target structural model parameters on power. For example, we found that higher values of the non-target regression coefficients led to lower power in some conditions (e.g., when the target effect size was high and the reliability of the measurement model was low) but higher power in other conditions (e.g., when the target effect size was low and the reliability of the measurement model was high). These findings highlight the importance of conducting a power analysis specific to one's own model and including plausible values of all parameters.

Finally, a comparison of the dotted and solid lines in Figure 3.3 reveals a large effect of the analysis method on power: Conducting a multiple regression on observed composite scores instead of latent variable SEM resulted in greater power to detect the target regression effect (see also Westfall & Yarkoni, 2016). This difference occurred despite the fact that the regression estimates were attenuated due to measurement error: That is, there was *greater* power to detect *smaller* observed effects in regression compared to larger disattenuated effects in SEM. This finding is consistent with research suggesting that structural model parameters are estimated with lower precision than regression parameters (Ledgerwood & Shrout, 2011; Savalei, 2019). This

difference was especially pronounced when scale reliability was low: For example, with 3 items and $\lambda = .3$ (CD = .23), power to detect $\beta_{YX} = .4$ for the composite score regression analysis reached over .80 at *N* = 750, but the same sample size conferred just .14 power for the latent variable SEM analysis.

Discussion

Our simulations demonstrated how power to detect a true effect of latent variable *X* on latent variable *Y* controlling for latent variables *W* and *Z* varies as a function of sample size, effect size, measurement reliability, and the value of other structural model parameters. Some of these factors (sample size, effect size, and measurement reliability) are predictable, in the sense that increasing any of these will necessarily increase power. The effect of other structural parameter values appears to be much harder to predict. When all these factors are considered together, it is very difficult to provide general rules of thumb that can meaningfully inform sample size planning.

These results also demonstrate how power can be affected by the decision to use latent variable SEM rather than composite-based observed variable regression. This discrepancy between latent variable and observed variable regression analyses highlights the need to conduct power analyses specific to SEM, because a simpler regression-based power analysis may give a very misleading estimate of the sample size that is required to attain sufficient power for SEM.

We caution against generalizing the specific numeric relations from these simulations to different structural models. In addition to the factors that we manipulated in this study, power is strongly affected by the structure of the model. For example, we conducted simulations with just one latent predictor instead of three (Figure S3.5 in the supplementary material displays the results of those simulations) and found higher rates of convergence overall, similar performance

across SEM and composite regression analyses at moderate to high item reliabilities, and greater power overall. Thus, observations from these simulations should not be treated as rules of thumb unless one is interested in exactly this simulated model with these specific model parameters. This lack of generalizability is precisely why it is crucial for researchers to conduct power analyses that are based on their particular model and plausible parameter values. In the next section, we discuss how to conduct power analysis to detect a target effect in SEM, and we introduce a Shiny web app that helps researchers do so.

Conducting Power Analysis to Detect a Target Effect in SEM

Power analysis to detect a target effect can be conducted either via analytic calculations or Monte Carlo simulations. The analytic approach, as is typically used for *t*-tests or multiple regressions and implemented in such software programs as G*Power (Faul, Erdfelder, Buchner, & Lang, 2009), relies on the asymptotic (large sample) properties of the test statistic to determine its expected distribution in finite samples. If the sampling distribution of the test statistic is known, determining power is simply a matter of computing the proportion of results in that sampling distribution that would lead to rejecting the null hypothesis of a parameter equaling zero. In SEM, this approach gives solutions that are accurate in very large samples but hold only approximately in smaller samples (Lai & Kelley, 2011). In small-to-moderate sized samples, discrepancies between estimated and asymptotic parameter standard errors can lead to analytic power estimates that do not reflect expected power in practice.

Because asymptotic power calculations can be misleading in small samples, we recommend a Monte Carlo simulation approach, in which many random data sets are drawn from a hypothetical true population model to mimic the selection of multiple random samples from the

population. Monte Carlo simulations can be used to calculate power to detect a target effect in SEM with the following four steps:

(1) Specify a hypothesized true population model and all of its parameter values;

(2) Generate a large number (e.g., 1000) of samples of size *N* from the hypothesized population model;

(3) Fit a structural equation model to each of the generated samples, recording whether the target parameter is significantly different than 0;

(4) Calculate power as the proportion of simulated samples that produce a statistically significant estimate of the parameter of interest.

This approach was popularized by Muthén and Muthén's (2002) guide on conducting Monte Carlo simulations to determine sample size in SEM studies using Mplus (see also Hancock and French, 2013). Yet this approach requires that users have access to Mplus (Muthén & Muthén, 1998–2017) and know how to program simulations using Mplus syntax and commands. To address the limitations of existing resources on power analysis to detect a target effect in SEM and to help researchers conduct their own, we introduce pwrSEM, a new Shiny web app. pwrSEM estimates power by running Monte Carlo simulations based on a model and sample size that users specify via a guided, step-by-step point-and-click interface. It accommodates a wide range of structural equation models, requires no experience in conducting simulations, and provides a suite of features that help researchers choose the model features that underlie their power analysis. Users can access pwrSEM online at: yilinandrewang.shinyapps.io/pwrSEM/. Alternatively, users can also run pwrSEM locally on their computer by downloading the R

source code file at https://osf.io/tcwyd/, opening it in R Studio, and then pressing the "Run App" link in the top-right hand corner of the R script section of the R Studio Window.[38]

A Tutorial on Power Analysis to Detect a Target Effect Using pwrSEM

To illustrate how to use pwrSEM, we present a scenario in which a researcher is interested in powering a study to detect an indirect effect in a mediation model. We will walk readers through each step of conducting the power analysis and describe the basic layout and functions of pwrSEM, both from the perspective of general users and from the perspective of the researcher. For simplicity's sake, we assume that the researcher in this scenario has a good sense of the population model and its parameter values. In the supplemental material, we consider a more complex scenario in which a researcher is interested in powering a study to detect several paths in a model, discuss some realistic challenges that users might confront when running a power analysis (e.g., specifying reasonable values of parameters, including factor loadings, structural paths, and residual variances), and highlight solutions that pwrSEM offers.

Research Scenario

Suppose that a researcher is interested in planning a study to test a simple mediation model. This model contains a predictor *X*, a dependent variable *Y*, and a mediator *M*, each modelled as a latent variable measured by three indicators. The regression coefficients from *X* to *M* and from *M* to *Y* are respectively labelled as *a* and *b* paths (see Figure 3.4 for a diagram of the model). The researcher would like to power their study to detect a true indirect effect of $a \times b$.

[38] The procedure on running a Shiny app locally is current as of R Studio version 1.2.5001.

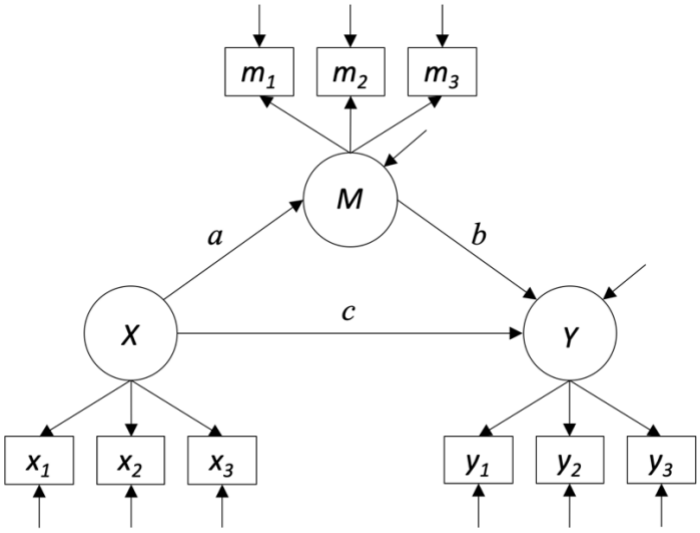

Figure 3.4. Mediation model used in the research scenario. All paths correspond to freely estimated parameters, except for latent variable (residual) variances, which were fixed to 1 for model identification.

Using pwrSEM to Conduct Power Analysis to Detect the Target Effect

Upon launching the app, users see the greeting screen (Figure 3.5). The left side panel provides a quick "how to" guide on using the app. The main panel on the right is where users run their power analysis, and it is divided into six tabs: "1. Specify Model," "2. Visualize," "3. Set Parameter Values," "4. Estimate Power," "Help," and "Resources." The first four tabs are ordered by the four steps that users take to conduct power analysis to detect a target effect; the "Help" and "Resources" tabs offer additional information that users might find helpful during the process (we discuss these two tabs in detail in the supplemental material).

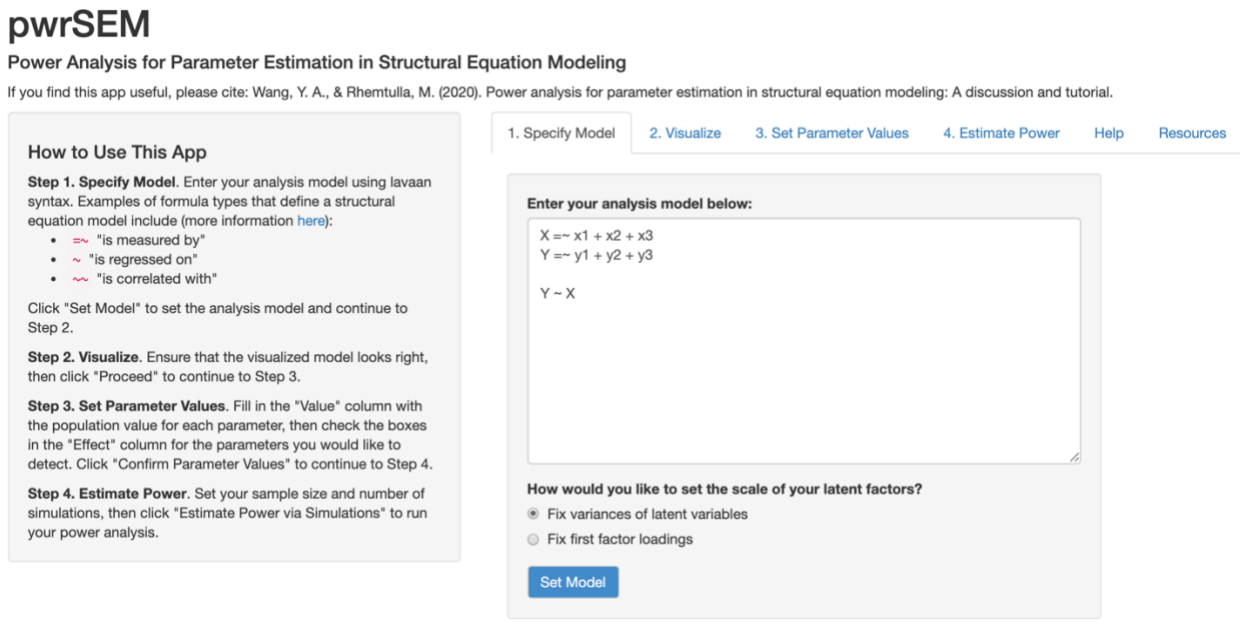

Figure 3.5. Greeting screen of pwrSEM. The box where users enter their analysis model is pre-filled with lavaan code of a sample model.

Step 1: Specify model. Users begin by specifying the structure of their analysis model. Currently, pwrSEM accepts lavaan syntax (Rosseel, 2012). After specifying a model, users can decide how they would like to set the scale of the latent variables. Selecting the default option will fix latent variable variances (or residual variances) to 1 and allow all factor loadings to be freely estimated. Alternatively, users can choose to fix the first factor loading to 1, allowing (residual) variances of latent variables to be freely estimated. Users confirm their model by clicking "Set Model," which will bring them to the next step.

In our scenario, the researcher specifies the measurement model (i.e., how latent variables *X*, *M*, *Y* are measured) and the structural paths among the constructs. Because the researcher is primarily interested in the indirect effect $a \times b$, they can label the component *a* and *b* paths by adding a* and b* in front of the corresponding predictors, then defining a new parameter *ab* as the product of the two paths (Figure 3.6). The researcher accepts the default option of fixing latent variables to unit variances and click "Set Model" to advance to Step 2.

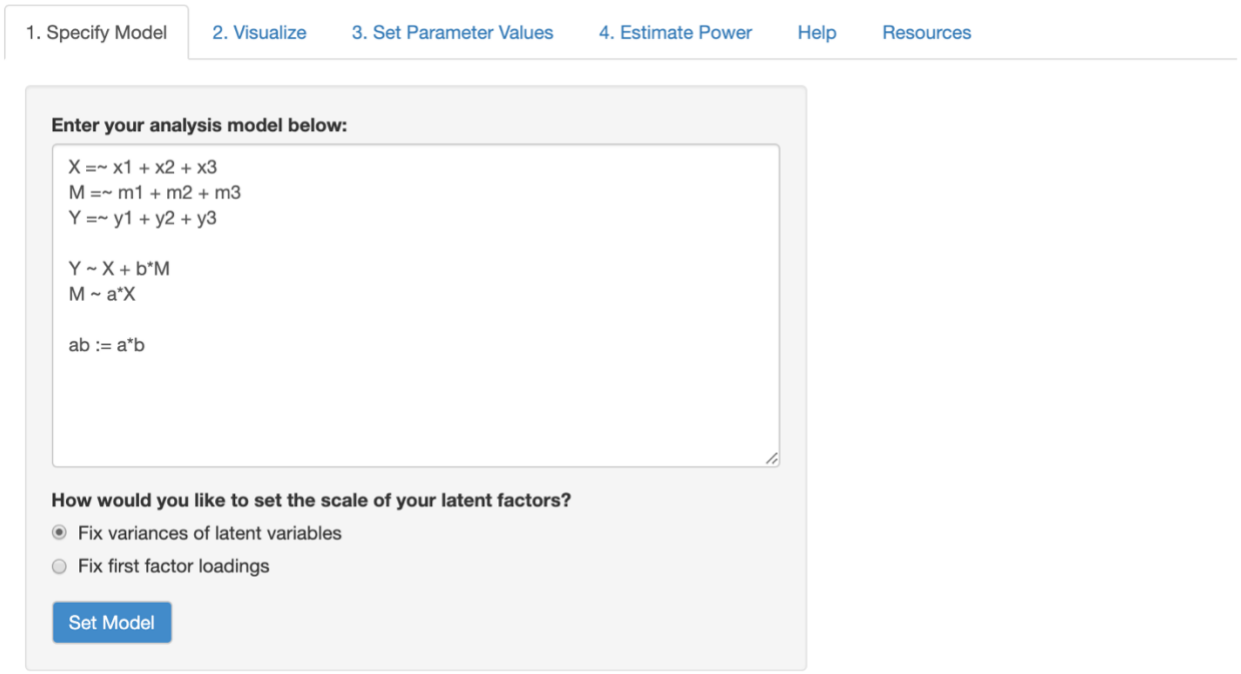

Figure 3.6. The researcher specifies their model in Step 1 of pwrSEM.

Step 2: Visualize. Upon proceeding to Step 2, users will see a path diagram of the model specified in Step 1 (generated by semPlot; Epskamp, 2015). Following SEM conventions, the diagram represents latent variables as circles, observed variables as squares, and linear regression coefficients as single-headed arrows. Double-headed loops that begin and end at the same variable represent variances (of exogenous variables) or residual variances (of indicators or endogenous variables), double-headed arrows connecting two variables represent covariances, and triangles represent means (of exogenous variables) or intercepts (of indicators or endogenous variables). Reflecting the decision in Step 1 to identify latent variables via fixing variances or factor loadings, fixed parameters are represented with dotted lines, and free parameters are represented with solid lines. Users can fine-tune the diagram with advanced visualization options, including the ability to change whether the measurement model is shown, change the size of shapes that represent observed and latent variables, and rotate the orientation of the

diagram. Once users visually confirm their model, they can click “Proceed” to continue to Step 3; otherwise, they can click “Back to Step 1” to modify the model.

In our scenario, the researcher visually confirms that their model is correct, then proceeds to Step 3 (Figure 3.7).

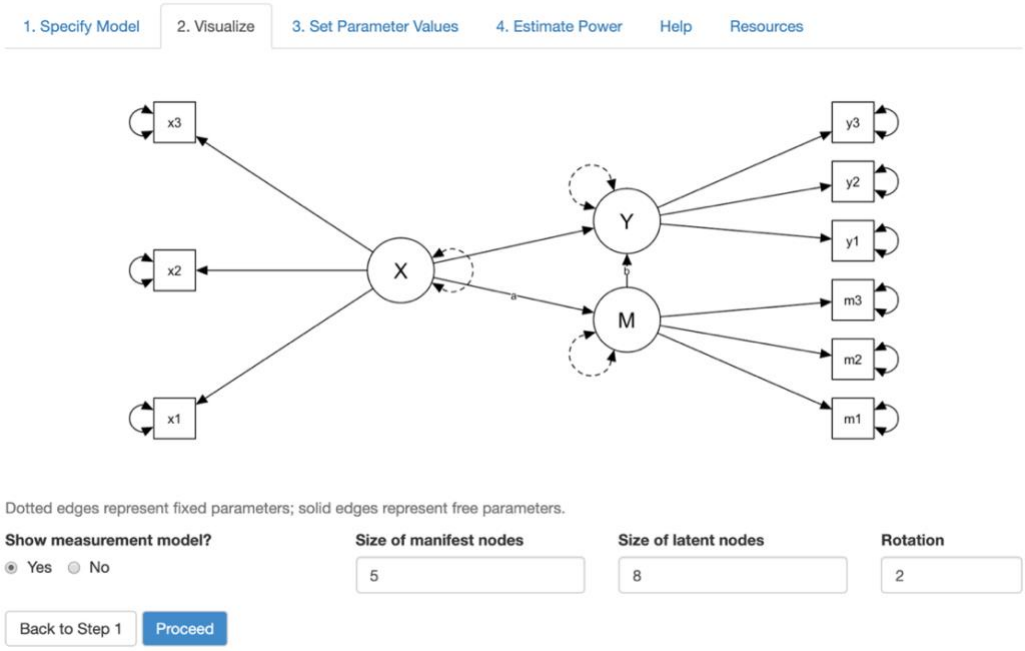

Figure 3.7. The researcher’s model as visualized in Step 2 of pwrSEM.

Step 3. Set parameter values. In Step 3, a list of all model parameters is automatically generated from the model set in Step 1 and placed in an interactive, editable table. The parameter table lists every model parameter, its user-specified label, description, type, and whether it is freely estimated. Users are prompted to set *all* population parameter values as listed in the “Value” column. Doing so is crucial because power to detect a target effect depends on both its value and the values of other parameters in the model: As illustrated by our simulations, power to detect a structural parameter of interest depends on the true size of that parameter as well as the number and true sizes of other parameters. Users set parameter values by double-clicking the corresponding cells in the table and entering their best estimate of the value of each parameter.

Note that the simulation procedure takes these estimates to be true population values; that is, it does not correct for the uncertainty of choosing parameter values. It is therefore important that users estimate power for a range of model parameter values to discover the sensitivity of the power estimate to such variation (we provide an example of a sensitivity analysis later in this tutorial and in the supplemental material). The table operates similarly to that in spreadsheet software (e.g., Excel): For example, users can copy-and-paste values from a spreadsheet directly into the table or copy the value of a cell to the cells below by dragging the bottom right corner of that first cell. Once all parameter values are set, users can select the target effects on which they would like to conduct power analysis and proceed to Step 4.

In our scenario, the researcher has a good idea of the likely population parameter values in their model and inputs those values into the parameter table. The researcher sets the factor loading of each indicator of *X* and *Y* to .70 (corresponding to a scale reliability of .74 for *X* and *Y*), the factor loading for each indicator of *M* to .80 (corresponding to a scale reliability of .84 for *M*), and the *a*, *b*, and *c* paths to .30, .20, and .10, respectively. Based on these values, the researcher calculates and sets the residual variance of each indicator of *X* to .51, the residual variance of each indicator of *M* to .36, the residual variance of each indicator of *Y* to .51, the total variance of *X* to be 1, and residual variances of *M* and *Y* to be .91 and .938, respectively. If the researcher does not know what values to set the residual variance parameters, they can enter the factor loadings and regression coefficients in the standardized metric, leave blank all other parameters, and click the button "Set Residual Variances for Me." pwrSEM will calculate and fill the residual variances (Figure 3.8); that is, the residual variances that are calculated will reflect the difference between a total variance of 1 and the variance that is accounted for by the entered model parameters. Lastly, the researcher sets the indirect effect of interest (*ab*) as the

product of the *a* and *b* paths, .06. Then the researcher checks the box in the "Effect" column that corresponds to their effect of interest and clicks "Confirm Parameter Values" to proceed to Step 4. In the supplemental material, we will discuss in further detail how the "Help" tab in pwrSEM can help users set parameter values if they do not have a good idea what those values should be.

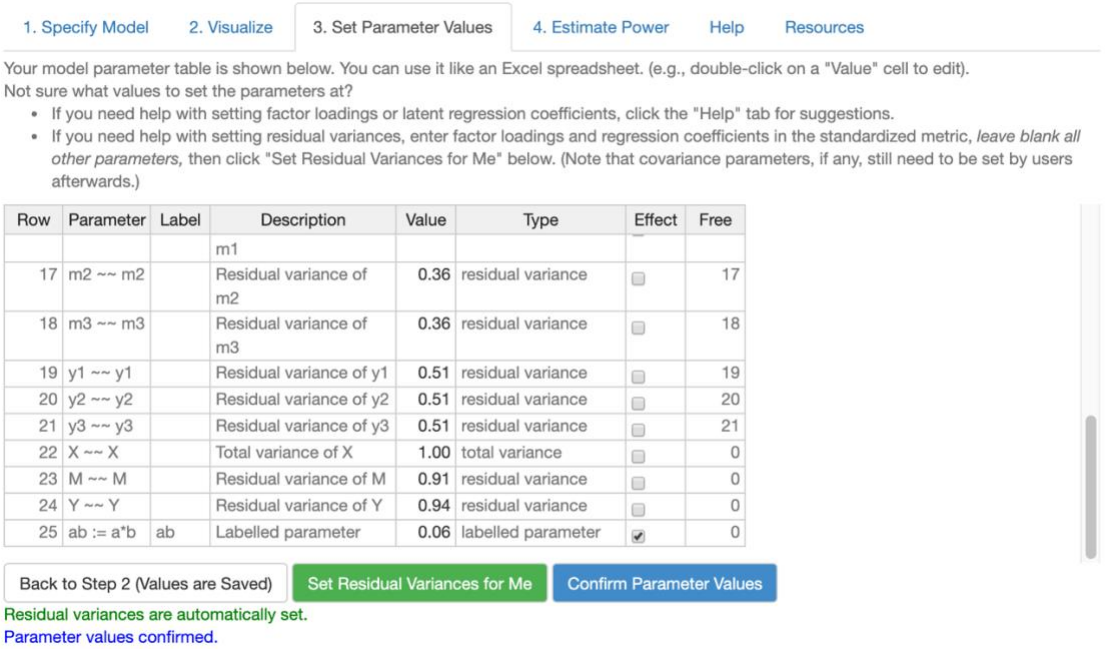

Figure 3.8. The researcher sets population parameter values in Step 3. They enter the factor loadings and regression coefficients in the standardized metric, click "Set Residual Variances for Me" to set the residual variances, and enter the population value of the indirect effect (the labelled parameter). Note that only some of the parameters are shown in this screenshot.

Step 4. Estimate power. The last step in power analysis is to choose a sample size and the number of samples to simulate. Users might initially specify a feasible sample size based on resources, power to detect misspecification, or other considerations. The number of samples to simulate reflects the desired precision of the power estimate, assuming no uncertainty of the population model: A larger number of samples returns a more precise power estimate but takes longer to run. We recommend that users start with a smaller number of samples (e.g., 100) to get a rough estimate of power before confirming it with a higher number. Optionally, users can also set desired alpha level and simulation seed (to get computationally reproducible results). Once

sample size and number of samples to simulate are set, users can click "Estimate Power via Simulations" to start running their power analysis. A progress bar will appear at the bottom right corner of the app interface to show users how many samples have been completed. Once simulations are complete, a power table and two histograms will appear. The power table shows each target effect and estimated power to detect it as the proportion of converged simulated samples with a statistically significant estimate of the target effect ("Power"), as well as a number of other outputs that users might find relevant, such as the convergence rate in the table note. Below the power table, users can find a histogram of the *p*-values and estimates of a given target effect from the simulated samples.

The researcher starts with a sample size of $N = 200$ and runs a simulation with 100 samples. Results suggest that they have .33 power to detect $ab = .06$ in their model. They increase their sample size to $N = 460$, and simulations suggest that they now have .81 power. The researcher confirms this result by re-running the simulation with 1,000 samples, which gives them a power estimate of around .85 (Figure 3.9). To explore the degree to which this power estimate is sensitive to the researcher's specifications of the population parameter values, the researcher re-runs the simulations to determine power for the target effect with $N = 460$ under 8 other sets of parameter values, modifying both *b* and the factor loadings of *M* (see Table 3.2 for details). This sensitivity analysis reveals that varying *b* and varying λ_M both have an impact on power. Thus, the researcher concludes that a sample of $N = 460$ will give them power of .85 to detect an indirect effect of .06 in their mediation model, but that their power might be lower if the reliability of *M* is lower or the size of *b* is smaller than specified.

Table 3.2

Power as a Function of the Population Value of b *and the Factor Loadings of* M

b	λ_M	Power
.15	.70	.53
	.80	.58
	.90	.65
.20	.70	.79
	.80	.85
	.90	.90
.25	.70	.95
	.80	.98
	.90	.98

Note: All power estimates were from running 1000 simulations with $N = 460$ using the same population model as described in the example, except for changes to *b* (and consequently *ab*) and λ_M. Residual variances were modified accordingly to maintain unit variances of the latent variables. Scale reliabilities of *M* were .74 ($\lambda_M = .70$), .84 ($\lambda_M = .80$), and .93 ($\lambda_M = .90$).

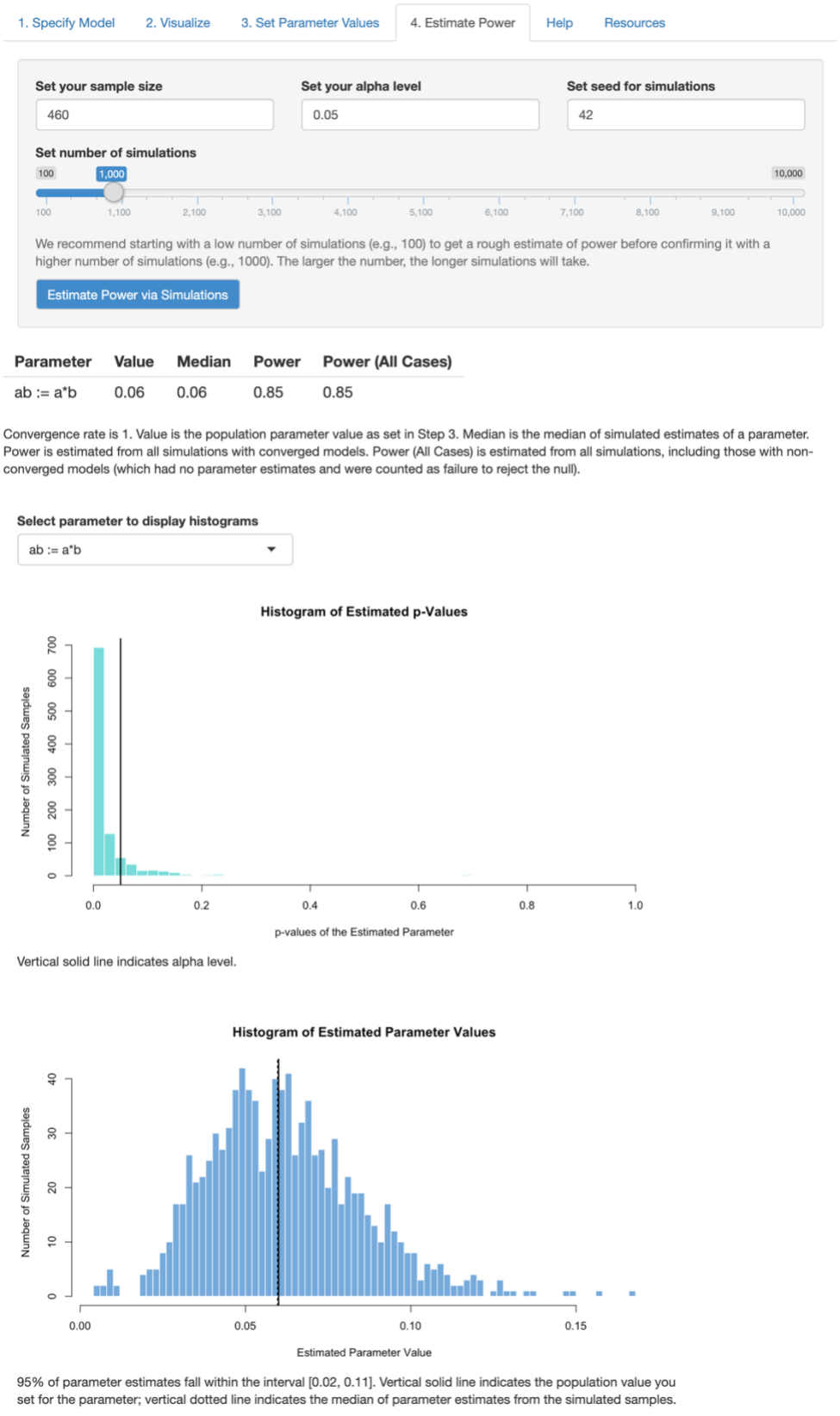

Figure 3.9. Results of the researcher's power analysis in Step 4.

Resources

The "Resources" tab offers additional resources that researchers might find useful. Although this tutorial and the app focuses on power to detect target effects, we emphasize that power to detect model misspecifications is also an important consideration. Thus, we included an additional calculator that allows researchers to run power analysis to detect model misspecification using the dominant approach based on *RMSEA* by MacCallum et al. (1996). We also provide additional learning resources for SEM for interested users.

Current Limitations and Potential Extensions

We acknowledge that the current version of pwrSEM has a number of limitations. First, to make the app more accessible and user-friendly, we assume that the population model from which the app generates data has the same parameters as the analysis model with which researchers plan to fit the data. In practice, this might not always be the case (e.g., researchers might intentionally choose a simpler analysis model). Second, because of the computational intensity of fitting structural equation models to a large number of simulated samples, the app currently does not allow for calculation of sample size based on desired level of power. Future work could address this limitation either by using a varying parameters approach (see Schoemann, Boulton, & Short, 2017; Schoemann, Miller, Pornprasertmanit, & Wu, 2014) or solving for *N* analytically and confirming the value via simulation. Third, the app currently only generates normally distributed data; future work on the app might be able to accommodate data with other distributional properties, such as categorical data, non-normal data, and data with a certain amount of missingness. Meanwhile, we encourage researchers interested in these more advanced specifications to implement them in R or other software environments directly.

We remind readers that SEM is a complex statistical technique. Although pwrSEM facilitates power analysis to detect a target effect in SEM, it assumes that researchers have basic working knowledge of conducting and interpreting SEM. We encourage researchers new to SEM to consult introductory learning resources for SEM (e.g., Kline, 2016; lavaan tutorial: http://lavaan.ugent.be/tutorial/). A list of such learning resources can be found under the "Resources" tab in the app.

Discussion

Power analysis in SEM can be challenging. This is especially true for power analysis to detect a target effect, which poses technical barriers for many researchers. Consequently, such power discussions remain scarce in the empirical SEM literature, and sample size planning based on rules of thumb is still common. This consequence is unfortunate in the current era of moving toward more robust research (Begley & Ellis, 2012; Ledgerwood, 2016; McNutt, 2014; Nosek, Spies, & Motyl, 2012; Nyhan, 2015; Vazire, 2017): As the field of psychology and many other sciences seek to improve statistical power and recognize the problems with underpowered studies, understanding what power is and seeking ways to increase it are important steps toward more accurate, reliable findings (Button et al., 2013; Cohen, 1962).

Of course, it should be emphasized that power is not the only consideration in research design, and sometimes different research design considerations may be at odds with each other. For example, our simulation study revealed that power in SEM is strongly affected by item reliability and scale length, but we would not encourage researchers to choose a more reliable but potentially less valid measure of their target construct for the sake of increasing power, without considering how such decisions might affect other aspects of research design. A researcher who replaces an existing measure with a more reliable one might end up inadvertently measuring a

narrower or altogether different construct. Not only would doing so compromise construct validity and limit the theoretical usefulness of the measure, but it would also change the population effect size to be detected (e.g., as in the case of "bloated specific" variables; Cattell, 1966). Researchers should balance their power goals with other desired ends (e.g., using resources efficiently, achieving estimation accuracy, maintaining procedural fidelity with past research), and tailor research design decisions to their specific research contexts (Finkel, Eastwick, & Reis, 2015; Ledgerwood, 2019; Maxwell, Kelley, & Rausch, 2008; Miller & Ulrich, 2016; Wang & Eastwick, in press; Wang, Sparks, Hess, Gonzales, & Ledgerwood, 2017).

Yet the case remains that if researchers take seriously what they can learn from their structural equation models, then they need to move beyond rules of thumb, evaluate the power implications of their models, and make planning and inferential decisions accordingly. By illustrating how power to detect a target effect in SEM is affected and introducing a new Shiny app for power analysis, we hope the current tutorial will help researchers develop informed understanding of power in SEM and allow them to incorporate power analysis into their empirical research pipeline.

Supplemental Material

Simulations: Power to Detect a Target Effect

Details of the Study Design and Method

We generated data consistent with a latent multiple regression in which three correlated latent predictors (*X, W, Z*) predicted a latent outcome *Y* (see Figure 3.1). We systematically varied three features of the population model, using a range of values that are typical in psychological research: (1) the population standardized effect size of the target effect, β_{YX}, ranging from 0.1 to 0.5 in increments of 0.1; (2) the value of all standardized factor loadings, ranging from λ = .3 to .9 in increments of .2 (for any given population model, all indicators had equal factor loadings); and (3) the number of indicators per latent factor, *p*/*f*, was either 3, 5, or 10, for a total of 12, 20, or 40 observed variables in the model. These parameter values allowed us to generate power curves that will show how target effect size (i.e., size of β_{YX}), item reliability (i.e., factor loadings), and number of items (i.e., scale length) produce differences in power to detect a target latent variable regression parameter. Each of these factors is expected to affect power, in both multiple regression and in SEM (e.g., Gerbing & Anderson, 1985; Williams et al., 1995).

In the main simulation study, the predictors were all intercorrelated at $\rho = .3$, and the regression coefficients of *W* and *Z* predicting *Y* were held constant at $\beta_{YW} = .1$ and $\beta_{YZ} = .2$. From each of the multivariate normal population distributions described by the study's 5 (effect size) × 4 (factor loading) × 3 (number of indicators per factor) = 60 populations, we drew 10,000 samples of each size *N*, which ranged from 50 to 1000 in increments of 50. We fit two models to each generated dataset (see Figure 3.2): The first was a structural equation model that corresponded to the population generating model (Figure 3.2 shows only the version with *p*/*f* =

3), with all factor loadings, factor covariances, indicator residual variances, and regression coefficients freely estimated. Latent variable (residual) variances were fixed to 1 to identify the models. The second was a multiple regression model based on composite scores formed by summing the indicators of each latent factor. For each fitted model, we recorded (a) whether the estimation algorithm successfully converged on a proper set of parameter estimates and was able to estimate standard errors, and (b) whether the *p* value associated with the target parameter, β_{YX}, was less than .05.

We also ran two additional sets of simulations: In Set 1, the predictors were all intercorrelated at $\rho = .5$, and the regression coefficients of *W* and *Z* predicting *Y* were held constant at $\beta_{YW} = .1$ and $\beta_{YZ} = .2$; in Set 2, the predictors were all intercorrelated at $\rho = .3$, and the regression coefficients of *W* and *Z* predicting *Y* were held constant at $\beta_{YW} = \beta_{YZ} = .3$. All other aspects of the population generation model were identical to that in the main simulation study. For each sample size *N*, we drew 1,000 samples.

In all simulations, we used R (R Core Team, 2019) and the R package *lavaan* (Rosseel, 2012) for data generation and analysis. The simulation code and results are available at https://osf.io/h8yfk/.

Results for Convergence Rates

Figure S3.1 shows convergence rates for structural equation models across all conditions when $\lambda = .3$ or .5 in the main simulation study (higher factor loadings resulted in near-perfect convergence rates and are not displayed). Consistent with previous literature, model convergence was affected by sample size, item reliability, and number of indicators per factor (Marsh et al., 1998). In particular, the convergence rate was greater than 90% in conditions with factor loadings .7 or higher, in conditions with 10 indicators per factor when $N \geq 150$, and in all

conditions when $N \geq 800$. Serious convergence problems arose when there were only 3 unreliable indicators per factor ($\lambda = .3$), resulting in convergence rates below 10% at N = 50 and below 90% at $N \leq 650$. The value of the target regression coefficient (β_{YX}) had little discernible effect on convergence. In the composite score multiple regression models, convergence rates were 100% in all conditions. Results for convergence rates in the two additional sets of simulations were nearly identical to the results in the main simulation study, so they are not reported here.

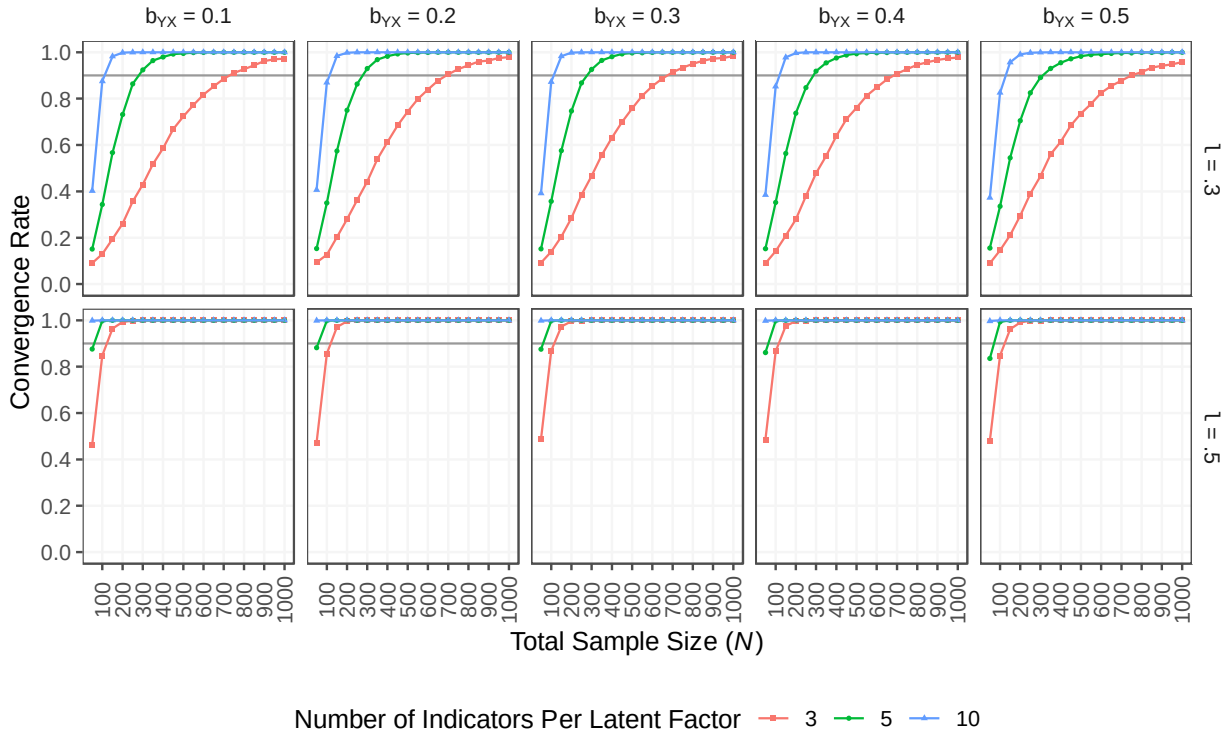

Figure S3.1. Convergence rate (y-axis) of structural equation models as a function of the total sample size (N; x-axis), population effect size (β_{YX}; columns), factor loadings (λ; rows), and number of indicators per latent factor (p/f; line color and point shape). Not shown here are conditions with λ = .7 or .9, all of which had convergence rates above 97%.

Additional Results

As described in the main text, we computed power as the proportion of converged cases in each condition in which the estimated target regression coefficient (β_{YX}) was significantly

different from 0 ($\alpha = .05$). To check if nonconvergence affected the pattern of results we report in the main text, we also examined results from calculating power as the proportion of all 10,000 simulations per condition in which β_{YX} was significantly different from 0 ($\alpha = .05$), where nonconverged models counted as not rejecting H_0. Results were largely similar (Figure S3.2).

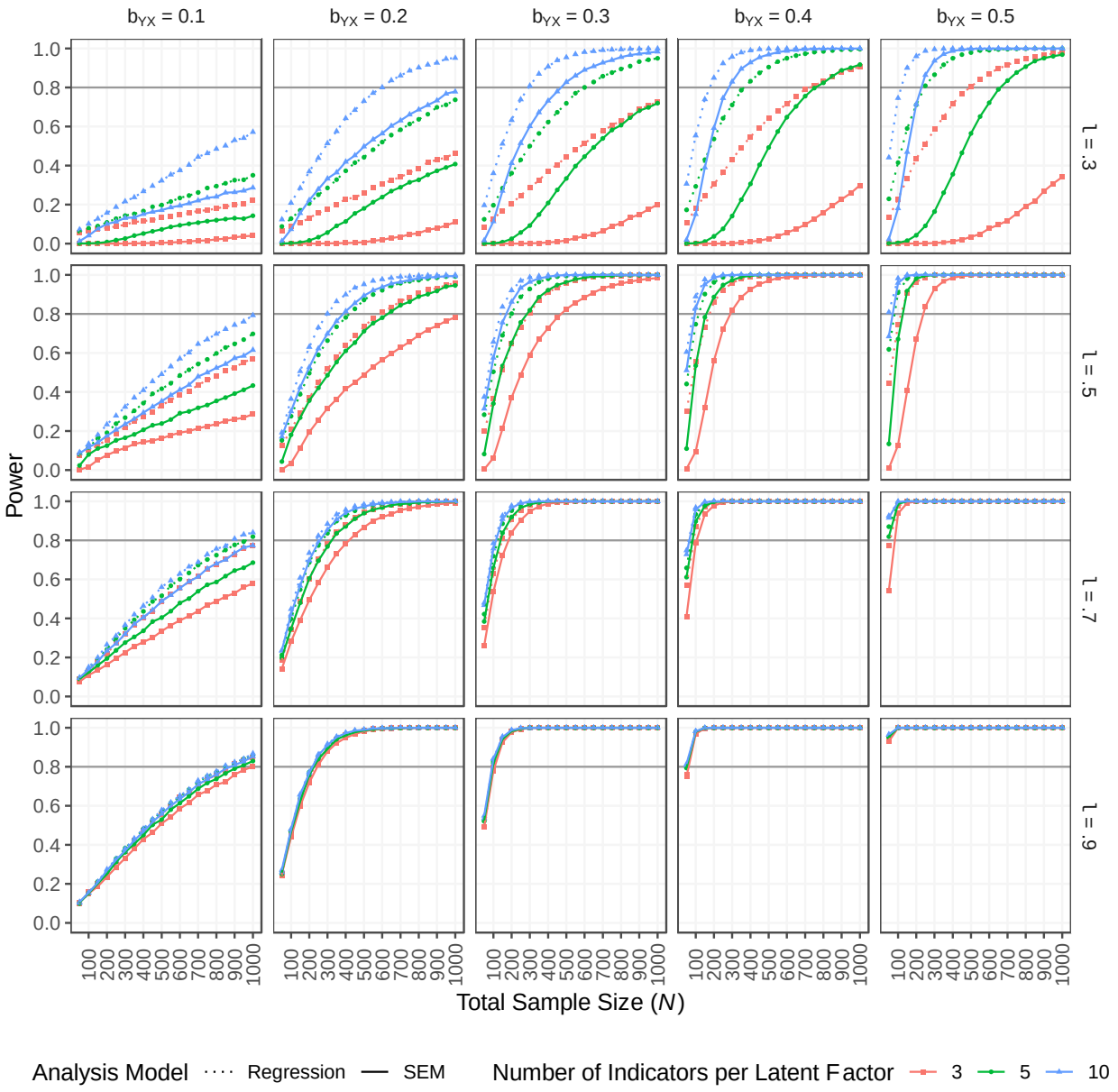

Figure S3.2. Power (y-axis) as a function of the total sample size (*N*; x-axis), population effect size (β_{YX}; columns), factor loadings (λ; rows), number of indicators per latent factor (line color and point shape), and analysis model (SEM or composite-score regression; line type). Power was

calculated as the percentage of all 10,000 simulations per condition in which β_{YX} was significantly different from 0 ($\alpha = .05$).

We also compared the results from our main simulation study to those from the two additional sets of simulations. Figures S3.3 and S3.4 show these comparisons. In both figures, solid lines are power curves for latent-variable SEM in the main simulation study (they exactly match the power curves for SEM in Figure 3.3, also in solid lines). Varying the intercorrelations among predictors (dotted lines in Figure S3.3) and varying the nontarget structural parameters (dotted lines in Figure S3.4) both produced discrepant results from the original set of simulations, suggesting that both factors affect power to detect the target effect.

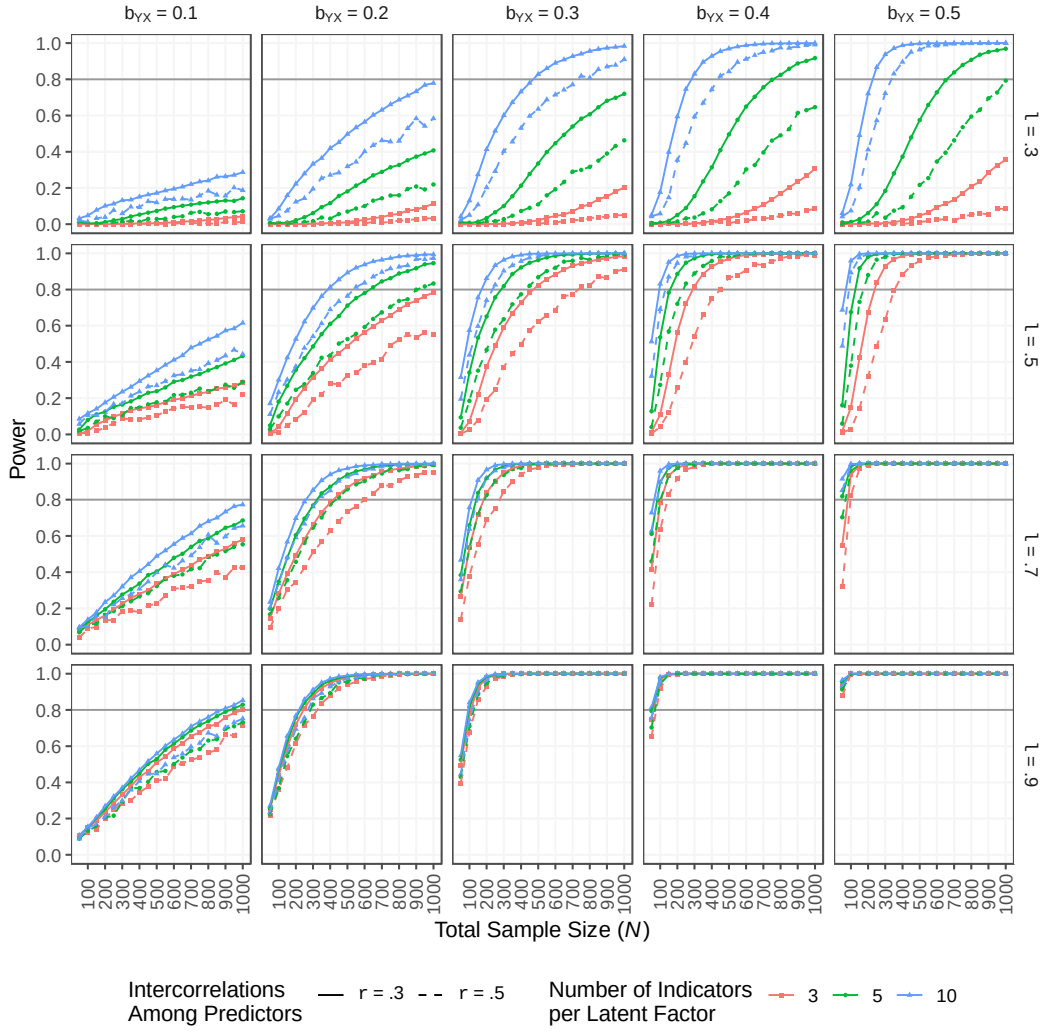

Figure S3.3. Power (y-axis) as a function of the total sample size (*N*; x-axis), population effect size (β_{YX}; columns), factor loadings (λ; rows), number of indicators per latent factor (line color and point shape), and intercorrelations among predictors (ρ; line type) in latent-variable SEM.

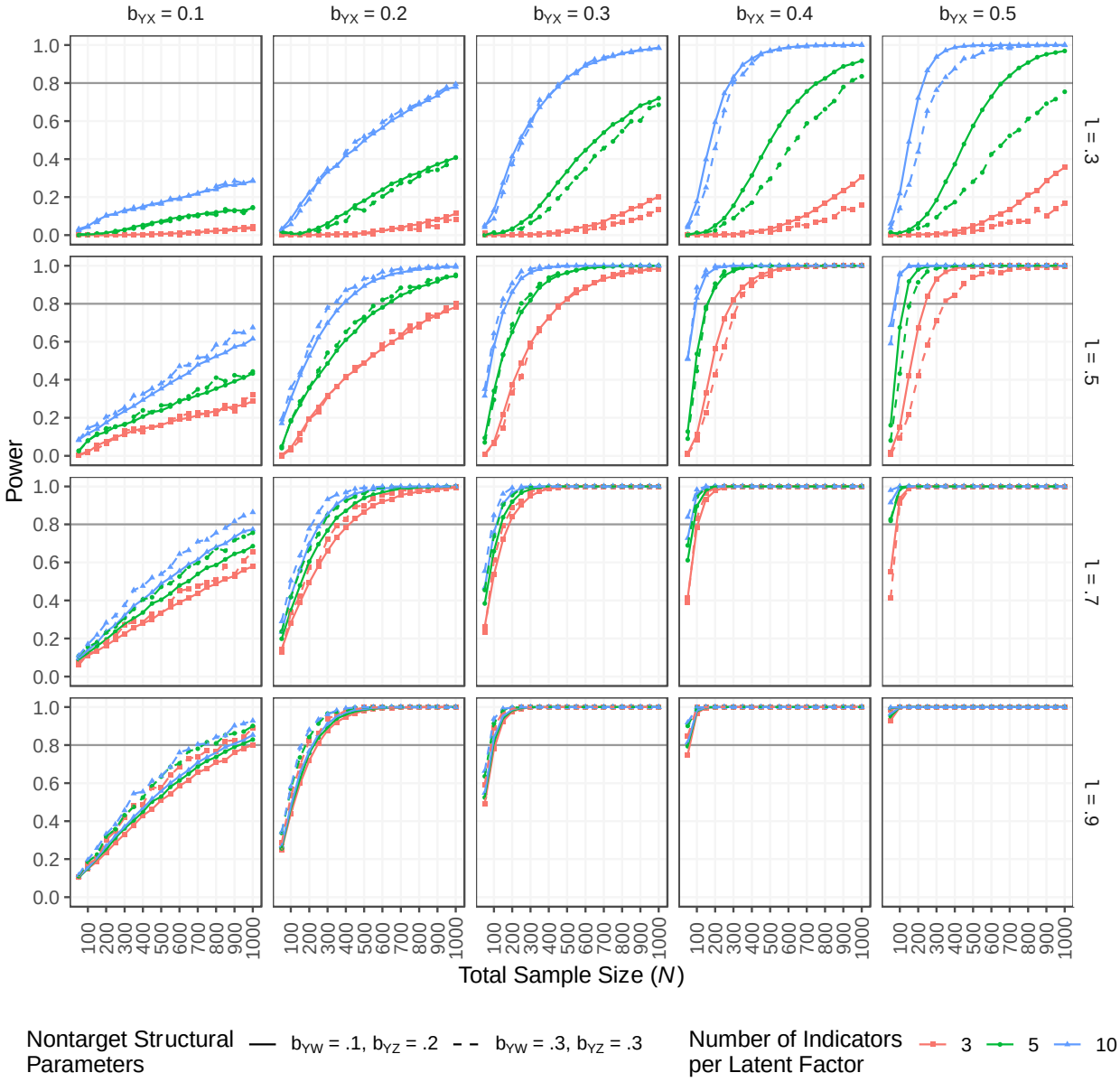

Figure S3.4. Power (y-axis) as a function of the total sample size (*N*; x-axis), population effect size (β_{YX}; columns), factor loadings (λ; rows), number of indicators per latent factor (line color and point shape), and nontarget structural parameters (β_{YW} and β_{YZ}; line type) in latent-variable SEM.

Lastly, we conducted simulations with only one latent predictor instead of three and found higher rates of convergence overall, similar performance across latent-variable SEM and

composite-score observed-variable regression analyses at moderate to high item reliabilities, and greater power overall (Figure S3.5).

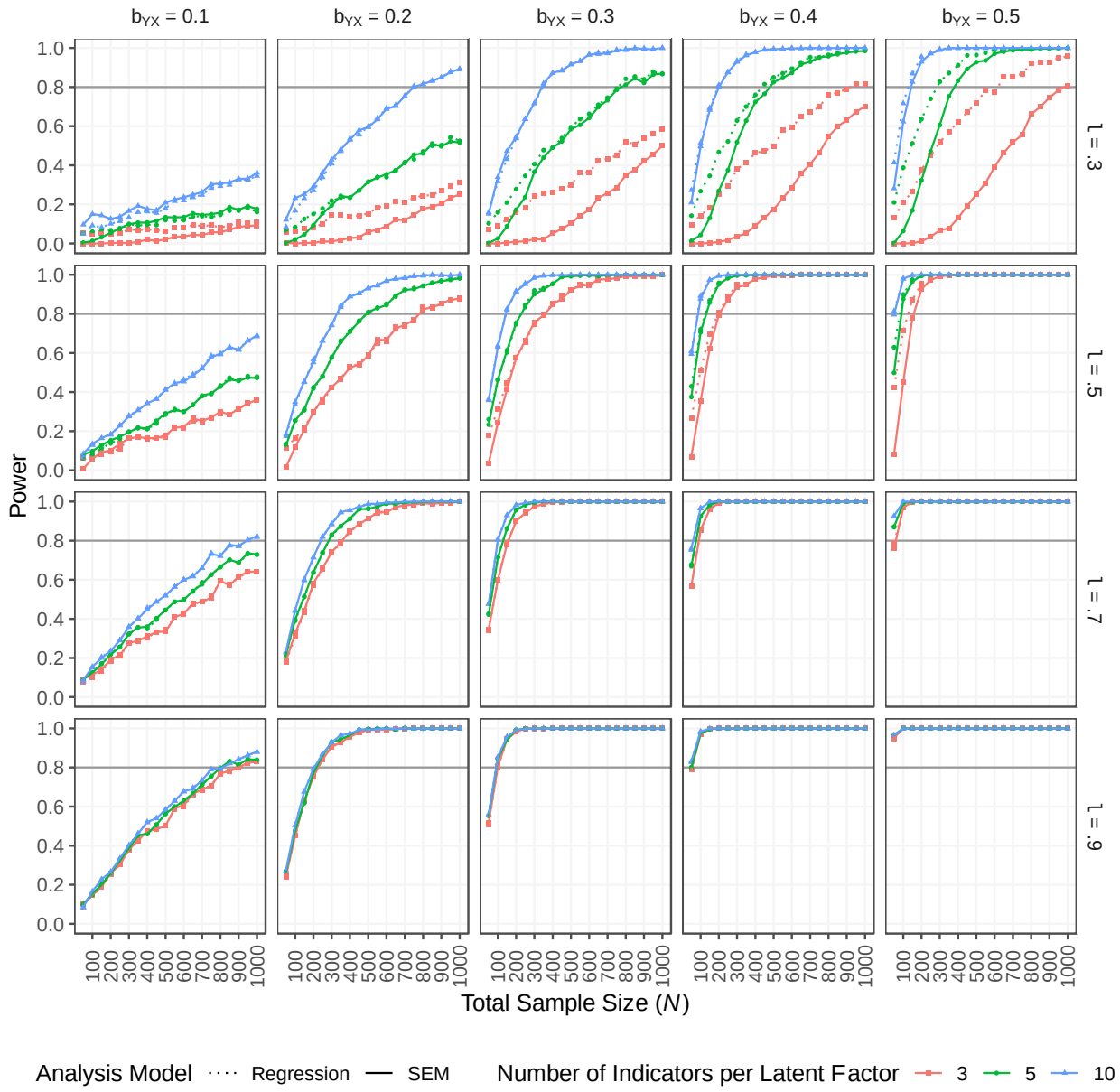

Figure S3.5. Power (y-axis) as a function of the total sample size (*N*; x-axis), population effect size (β_{YX}; columns), factor loadings (λ; rows), number of indicators per latent factor (line color and point shape), and analysis model (SEM or composite-score regression; line type). For each sample size N, we drew 1,000 samples. Power was calculated as the percentage of converged cases in each condition in which β_{YX} was significantly different from 0 (α = .05).

Conducting Power Analysis for Detecting a Target Effect in a Complex Model

In the main text, we presented a scenario in which the researcher has a good idea of the population model when conducting power analysis for detecting a target effect in SEM. In practice, of course, this is not usually the case. Researchers might confront realistic challenges such as specifying reasonable population values of parameters (e.g. factor loadings, structural paths, and residual variances) and examining how sensitive power estimates are to parameter value specifications. We highlight and provide solutions to these challenges below.

Research Scenario

We consider a complex hypothetical scenario in which a researcher, Professor Q, is interested in testing the effects of students' self-motivation on academic achievement, based on a widely cited model proposed by Zimmerman et al. (1992). Professor Q hypothesizes that students' current academic performance is predicted by their self-efficacy for academic achievement (SAA) and their grade goals (SGG), as well as three other indirect predictors, self-efficacy for self-regulated learning (SSL), parent grade goals (PGG), and prior academic performance (PAP; see Figure S3.6 for a diagram of the hypothesized structural model). Professor Q is primarily interested in the effect of SAA on CAP ($\beta_{CAP.SAA}$) and plans to conduct a power analysis for detecting that effect in a structural equation model.

Using *pwrSEM* to Conduct Power Analysis for Detecting the Target Effect

Step 1: Specify model. Professor Q plans to measure each construct in the same way that Zimmerman et al. (1992) did (see Table S3.1 for a summary) and specifies the measurement model accordingly in *pwrSEM*. Professor Q then specifies the structural paths among the constructs according to Figure S3.6. Professor Q accepts the default option of fixing latent variables to unit variances (Figure S3.7).

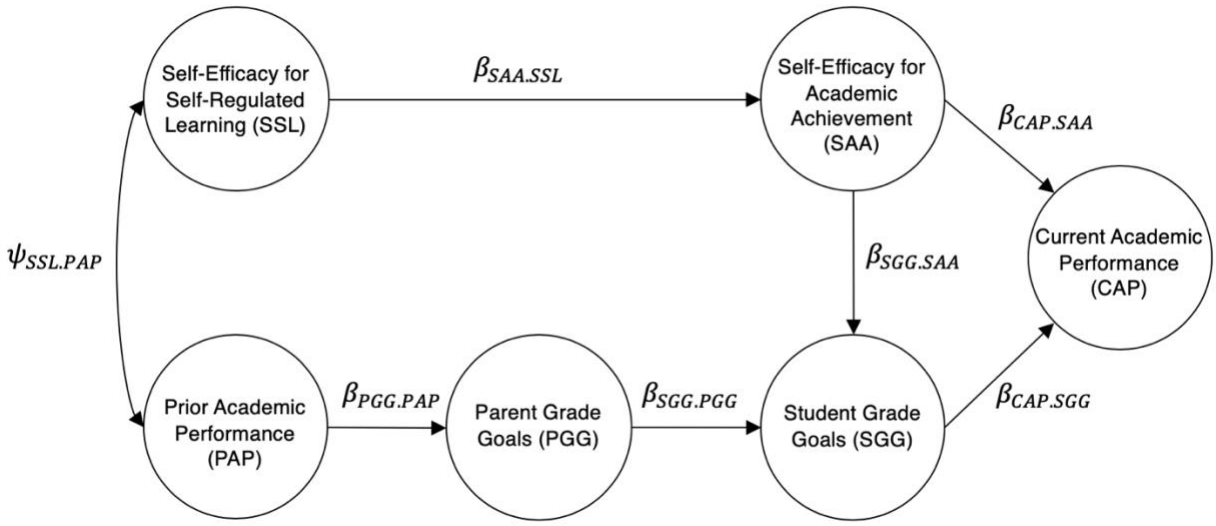

Figure S3.6. Professor Q's hypothesized structural model. For simplicity of presentation, latent variable (residual) variances are not shown. Structural regression coefficients are denoted in β, and the covariance between exogenous latent variables is denoted as ψ.

Table S3.1

Measures Used by Professor Q.

Construct	Measure	Measure Reference
Self-efficacy for self-regulated learning (SSL)	11 items measuring students' self-assessed ability to use a variety of self-regulated learning strategies (α = .87)	Bandura (1989)
Self-efficacy for academic achievement (SAA)	9 items measuring students' self-assessed ability to achieve in nine subject domains (α = .70)	Bandura (1989)
Students' grade goals (SGG)	2 items measuring students' expected grade and grade that students regard as minimally satisfying (α = .80)	Locke & Bryan (1968)
Parents' grade goals (PGG)	2 items measuring parents' expected grade and grade that parents regard as minimally satisfying (α = .63)	Locke & Bryan (1968)
Prior academic performance (PAP)	3 exam scores from prior course	-
Current academic performance (CAP)	3 exam scores in the current course	-

Note: Cronbach's alphas were reported in Zimmerman et al. (1992).

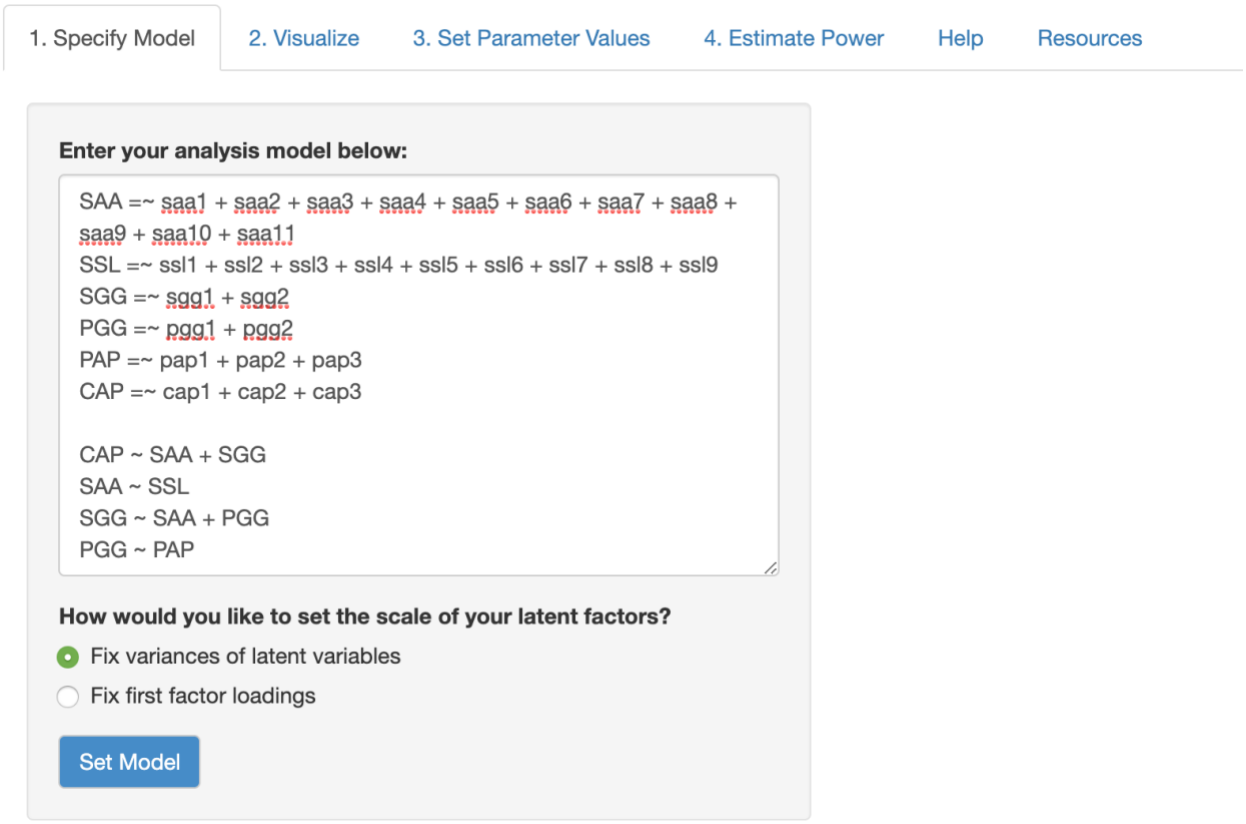

Figure S3.7. Professor Q specifies the model in Step 1 of *pwrSEM*.

Step 2: Visualize. Professor Q sees their model visualized in Step 2 (Figure S3.8). Because the model contains many observed variables, Professor Q could shrink the node size of the variables to see the diagram more clearly or simplify the diagram by choosing not to show the measurement model.

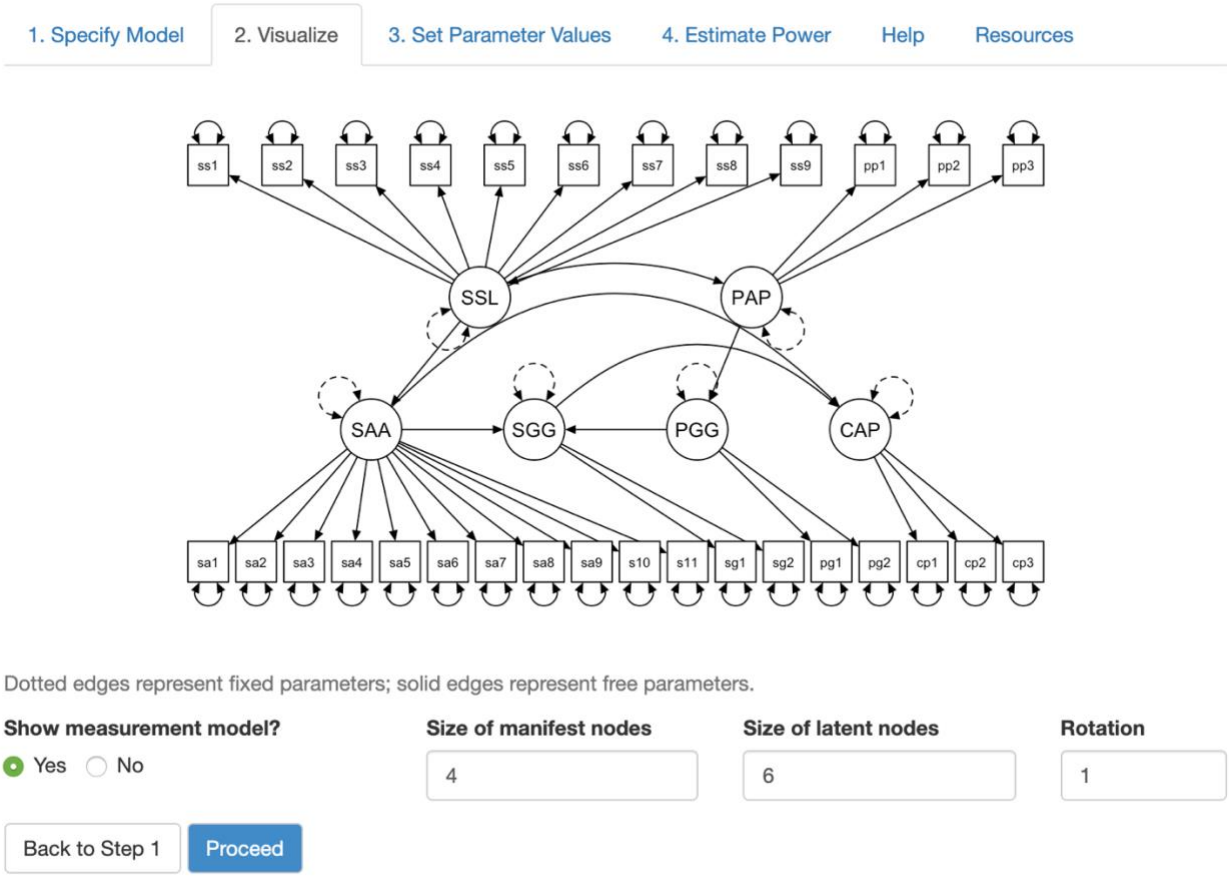

Figure S3.8. Professor Q's model as visualized in Step 2 of *pwrSEM*.

Step 3. Set parameter values. In the research scenario described in the main text, the researcher has a good sense of the population parameter values and directly enters them into the app. We suspect that realistically, a more likely scenario is that researchers only have a vague idea (or sometimes no idea at all) of the population parameter values and may find setting those values daunting. This scenario is normal: If the population values were known, there would be no need to do the research. Ideally, users can refer to existing literature for the likely values of the parameters. For example, if the model includes a latent variable measured by a well-established scale, users may be able to specify the factor loadings of the items in that scale by referring to estimates in prior work. Similarly, meta-analyses or large datasets on a relevant effect might be able to provide a point estimate of an effect size. In practice, however, such

information might not be readily available (e.g., due to lack of measurement work on certain constructs), directly applicable (e.g., when effect size estimates in the literature are based on observed variables, not latent variables), or accurate (e.g., due to publication bias).

Two features of *pwrSEM* address these challenges of setting parameter values. First, the "Help" tab contains calculators that help users set parameter values based on information they have. If information on factor loadings for a latent variable is not directly available but users know the number of indicators (e.g., number of items on a scale) and have an estimate of the overall scale reliability (e.g., Cronbach's alpha), they can calculate average factor loadings with the factor loadings calculator, which implements the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910; see Hancock & French, 2013, for a similar approach). Similarly, if information on structural parameters is not directly available but users have an estimate from observed variables, they can use the latent effect size calculator to calculate disattenuated structural parameters using Spearman's correction (1904).[39] Second, if users have standardized factor loadings and standardized regression coefficients but do not know what values to set for the residual variance parameters, they can enter all factor loadings and all regression coefficients into the parameter table, leave blank the other parameters, and click on "Set Residual Variances for Me" below the parameter table. The residual variances will be calculated and automatically filled. The calculated residual variances will reflect the difference between a total variance of 1 and the variance that is accounted for by the entered model parameters.

In our hypothetical scenario, Professor Q does not have direct information on what factor loadings to set, so they click on the "Help" tab and use the factor loading calculator to set the

[39] Note, however, that Spearman's correction might be inaccurate if the effect sizes of observed variables are estimated from a path model with more than two variables, because measurement error in path models can result in complicated patterns of over- and under-estimation of structural parameters (Cole & Preacher, 2014).

average factor loadings for four of the six latent variables, based on reliability estimates reported in Zimmerman et al. (1992): $\lambda = .62$ for SSL ($\alpha = .87$, 11 items), $\lambda = .45$ for SAA ($\alpha = .70$, 9 items), $\lambda = .82$ for SGG ($\alpha = .80$, 2 items), and $\lambda = .68$ for PGG ($\alpha = .63$, 2 items). Professor Q presumably has some information on how reliably their students perform academically; if not, Professor Q could still use the factor loading calculator to set factor loadings to plausible values (see Savalei, 2019). In the current scenario, Professor Q sets $\lambda = .76$ for both PAP and CAP ($\alpha = .80$, 3 items).

Because the path model reported in Zimmerman et al. estimated relations among observed variables, rather than latent variables, Professor Q chooses to directly set the following structural parameters based on Professor Q's best estimates: $\beta_{CAP.SGG} = 0.40$, $\beta_{SAA.SSL} = 0.50$, $\beta_{SGG.SAA} = 0.40$, $\beta_{SGG.PGG} = 0.40$, $\beta_{PGG.PAP} = 0.20$. They set the target effect size as $\beta_{CAP.SAA} = 0.20$, which is the smallest effect size they would be interested in detecting (Albers & Lakens, 2018). Alternatively, they could choose an effect size estimate from the literature. Note that due to potential publication bias, they might consider adjusting the effect size estimate downward to account for publication bias (Anderson et al., 2017), or use a more conservative estimate (e.g., a lower-bound estimate of the effect size; Perugini et al., 2014). Professor Q then uses the latent effect size calculator to set the correlation between the two exogenous latent variables (i.e., SSL and PAP) as $\psi_{SSL.PAP} = .17$, based on the point estimate of the correlation between SSL and PAP as composite-score observed variables ($r = .14$, Zimmerman et al., 1992) and the reliability of those two variables. Because the parameter values Professor Q enters are standardized values, they click on "Set Residual Variances for Me" to let *pwrSEM* autofill the residual variances in the model (Figure S3.9; see Table S3.2 for the full parameter table). Professor Q is interested in

powering their study to detect $\beta_{CAP.SAA}$, so they check the box in the "Effect" column for that parameter and click on "Confirm Parameter Values" to proceed to Step 4.

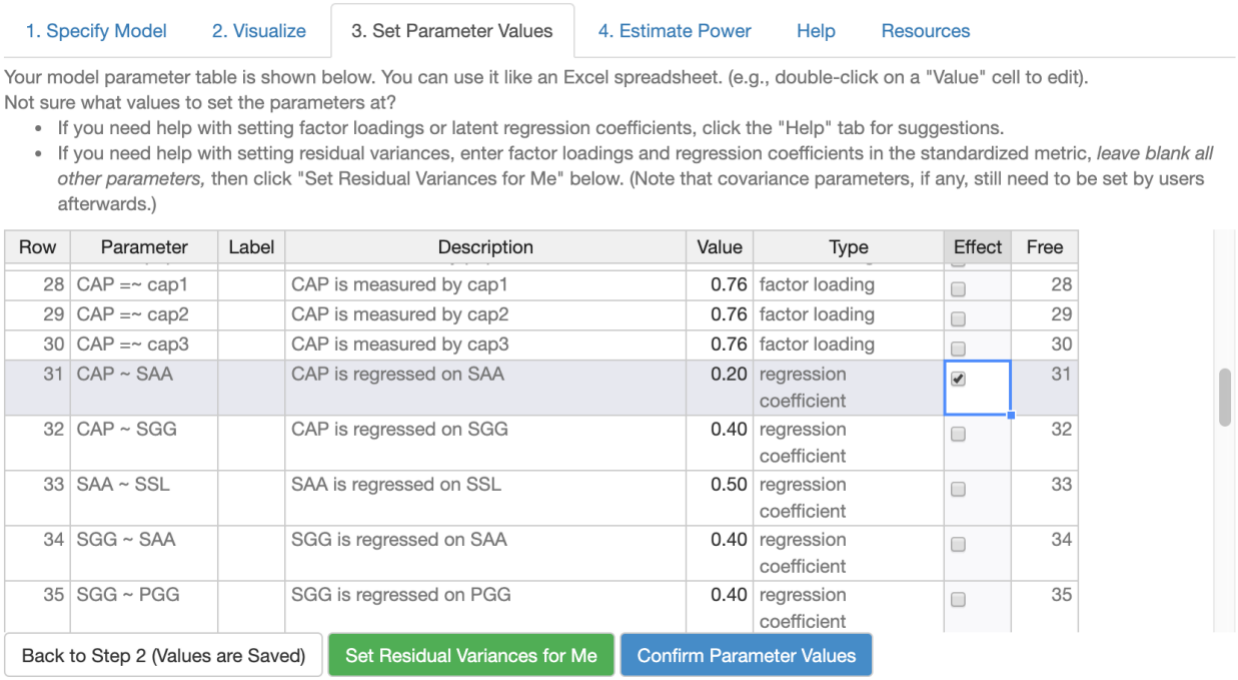

Figure S3.9. Professor Q sets population parameter values in Step 3. Note that only some of the parameters are shown in this screenshot.

Table S3.2

Parameter Table filled by Professor Q in Step 3 of pwrSEM.

Row	Parameter	Description	Value
1–11 of 73	SAA =~ saa1 … SAA =~ saa11	SAA is measured by saa1–saa11	0.45
12–20 of 73	SSL =~ ssl1 … SSL =~ss19	SSL is measured by ssl1–ssl9	0.62
21–22 of 73	SGG =~ sgg1 SGG =~ sgg2	SGG is measured by sgg1 and sgg2	0.82
23–24 of 73	PGG =~ pgg1 PGG =~ pgg2	PGG is measured by pgg1 and pgg2	0.68
25–27 of 73	PAP =~ pap1 … PAP =~ pap3	PAP is measured by pap1–pap3	0.76
28–30 of 73	CAP =~ cap1	CAP is measured by cap1–cap3	0.76

	…		
	CAP =~ cap3		
31 of 73	CAP ~ SAA	CAP is regressed on SAA	0.20
32 of 73	CAP ~ SGG	CAP is regressed on SGG	0.40
33 of 73	SAA ~ SSL	SAA is regressed on SSL	0.50
34 of 73	SGG ~ SAA	SGG is regressed on SAA	0.40
35 of 73	SGG ~ PGG	SGG is regressed on PGG	0.40
36 of 73	PGG ~ PAP	PGG is regressed on PAP	0.20
37–47 of 73	saa1 ~~ saa1	Residual variance of saa1–saa11	0.80
	…		
	saa11 ~~ saa11		
48–56 of 73	ssl1 ~~ ssl1	Residual variance of ssl1–ssl9	0.62
	…		
	ssl1 ~~ ssl9		
57–58 of 73	sgg1 ~~ sgg1	Residual variance of sgg1 and sgg2	0.33
	sgg2 ~~ sgg2		
59–60 of 73	pgg1 ~~ pgg1	Residual variance of pgg1 and pgg2	0.54
	pgg2 ~~ pgg2		
61–63 of 73	pap1 ~~ pap1	Residual variance of pap1–pap3	0.42
	…		
	pap3 ~~ pap3		
64–66 of 73	cap1 ~~ cap1	Residual variance of cap1–cap3	0.42
	…		
	cap3 ~~ cap3		
67 of 73	SAA ~~ SAA	Residual variance of SAA	0.75
68 of 73	SSL ~~ SSL	Total variance of SSL	1.00
69 of 73	SGG ~~ SGG	Residual variance of SGG	0.68
70 of 73	PGG ~~ PGG	Residual variance of PGG	0.96
71 of 73	PAP ~~ PAP	Total variance of PAP	1.00
72 of 73	CAP ~~ CAP	Residual variance of CAP	0.74
73 of 73	SSL ~~ PAP	Variance of SSL covaries with variance of PAP	0.17

Note. Rows with repeated factor loadings and residual variances are omitted. Values are rounded to two decimal points. Some columns shown in the app are omitted in this table.

Step 4. Estimate power. Professor Q starts with a sample size of 200 and runs a power analysis with 100 simulated samples. Results suggest that they have .61 power to detect $\beta_{CAP.SAA}$ = 0.20 in the model. They increase the sample size to 350, and simulations suggest that they now have .85 power. Professor Q confirms this result by rerunning the power analysis with 1,000

simulated samples, which gives a power estimate of .81 (Figure S3.10). Professor Q thus concludes that $N = 350$ will provide .81 power to detect the effect of SAA on CAP in the model.

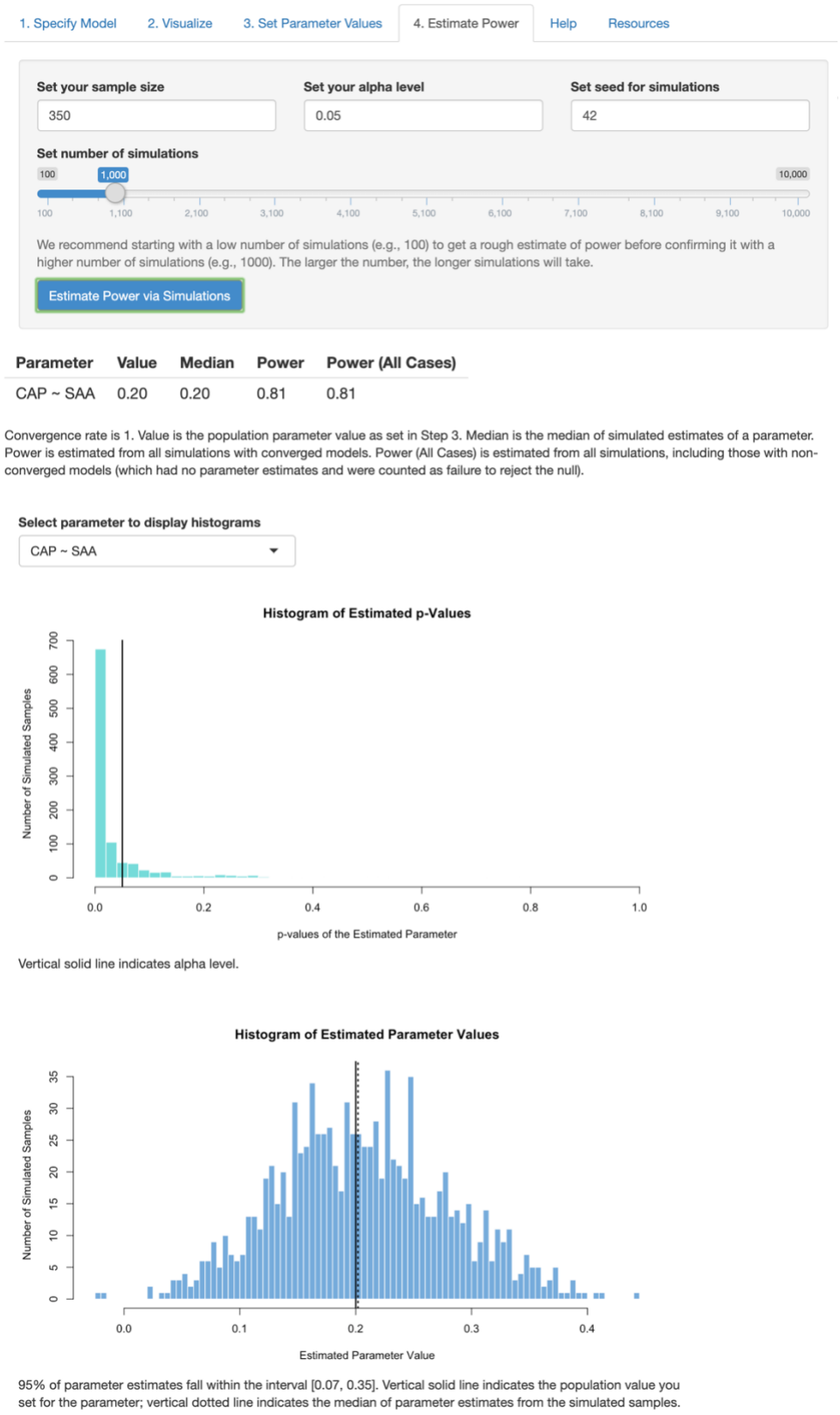

Figure S3.10. Results of Professor Q's power analysis in Step 4 of *pwrSEM*.

Given the uncertainty in setting population parameter values (Step 3), users may want to explore the degree to which power is sensitive to specifications of the population parameter values by varying them across a range of plausible values (Cook, 1986). Indeed, we recommend doing so to be more confident about the power estimates, especially when users are unsure of the population parameter values they specified. In Professor Q's scenario, they run a sensitivity analysis to test how robust results from their power analysis are to departures from the model parameters they initially specified. Professor Q reruns the power analysis for detecting the target effect with $N = 350$ under 8 other sets of parameter values, modifying both the factor loadings of the academic performance variables (i.e., PAP and CAP) and the size of the target effect (see Table S3.3 for details). This sensitivity analysis reveals that, as expected, modifying the target effect has a substantial impact on power, such that an effect size 25% smaller (i.e., $\beta_{CAP.SAA} = 0.15$) resulted in lower power at .55. In comparison, changes in the factor loadings of PAP and CAP have relatively minor impact on power. Given that $\beta_{CAP.SAA} = 0.20$ is the smallest effect size of interest, Professor Q concludes that 350 is a reasonable sample size.

Table S3.3

Power as a Function of the Population Value of the Target Effect and the Factor Loadings of the Academic Performance Variables in Professor Q's Model.

λ_{PAP}, λ_{CAP}	Power
	$\beta_{CAP.SAA} = 0.15$
.66	.49
.76	.55
.87	.61
	$\beta_{CAP.SAA} = 0.20$
.66	.72
.76	.81
.87	.86
	$\beta_{CAP.SAA} = 0.25$
.66	.91
.76	.95
.87	.96

Note: All power estimates were obtained from power analyses with 1,000 simulations of N = 350. The population model was specified as described in the scenario, except for changes to $\beta_{CAP.SAA}$, λ_{PAP}, and λ_{CAP}. Residual variances were modified accordingly to maintain unit variances of the latent variables. Scale reliabilities of PAP and CAP were .70 ($\lambda_{PAP} = \lambda_{CAP} = .66$), .80 ($\lambda_{PAP} = \lambda_{CAP} = .76$), and .90 ($\lambda_{PAP} = \lambda_{CAP} = .87$).

References

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*, 751-763.

Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T. M., Rosenberg, M. J., & Tannenbaum, P. H. (1968). *Theories of cognitive consistency: A sourcebook.* Chicago, IL: Rand McNally.

Aggarwal, P., Castleberry, S. B., Ridnour, R., & Shepherd, C. D. (2005). Salesperson empathy and listening: Impact on relationship outcomes. *Journal of Marketing Theory and Practice, 13*, 16-31.

Ahluwalia, R., & Gurhan-Canli, Z. (2000). The effects of extensions on the family brand name: an accessibility-diagnosticity perspective. *Journal of Consumer Research*, 27, 371-381.

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84*, 888-918.

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracın, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173-221). Mahwah, NJ: Lawrence Erlbaum.

Alcser, A., Smith, L. K., & Eastwick, P. W. (2021). Inferring one's own attitude towards a novel attribute: The moderating role of complexity in juice tasting. *Manuscript under review.*

Algoe, S. B., Dwyer, P. C., Younge, A., & Oveis, C. (2019). A new perspective on the social functions of emotions: Gratitude and the witnessing effect. *Journal of Personality and Social Psychology.* Advance online publication.

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Psychology/Revue canadienne de psychologie, 34*, 1–11.

Allan, L. G., Siegel, S., & Tangen, J. M. (2005). A signal detection analysis of contingency data. *Learning & Behavior, 33,* 250–263.

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology, 9*, 272–279.

Anderson, N. H. (1971). Integration theory and attitude change. *Psychological Review, 78*, 171-206.

Apostolou, M. (2007). Sexual selection under parental choice: The role of parents in the evolution of human mating. *Evolution and Human Behavior, 28*, 403-409.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258-290.

Bagozzi, R. P. (1980). *Causal models in marketing*. New York: John Wiley.

Baron-Cohen, S. (2011). *The science of evil: On empathy and the origins of cruelty*. NY: Basic Books.

Batson, C. D. (2009). These things called empathy: Eight related but distinct phenomena. In J. Decety & W. Ickes (Eds.), *The social neuroscience of empathy* (pp. 3–16). Cambridge, MA: MIT Press.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*, 531-533.

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*, 78-117.

Berndsen, M., Wenzel, M., Thomas, E. F., & Noske, B. (2018). I feel you feel what I feel: Perceived perspective-taking promotes victims' conciliatory attitudes because of inferred emotions in the offender. *European Journal of Social Psychology, 48*, 103-120.

Blanco, F. (2017). Positive and negative implications of the causal illusion. *Consciousness and Cognition, 50,* 56-68.

Blanco, F., Matute, H., & Vadillo, M. A. (2013). Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learning & Behavior, 41,* 333-340.

Bloom, P. (2017). Empathy and its discontents. *Trends in Cognitive Sciences, 21,* 24-31.

Bollen, K. A. (1989). Structural equations with latent variables. John Wiley & Sons.

Boomsma, A. (1982). Robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Part I, pp. 149–173). Amsterdam, the Netherlands: Elsevier.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 50,* 229-242.

Borgia, G. (1995). Complex male display and female choice in the spotted bowerbird: specialized functions for different bower decorations. *Animal Behaviour, 49,* 1291-1301.

Brambilla, M., & Leach, C. W. (2014). On the importance of being moral: The distinctive role of morality in social judgment. *Social Cognition, 32,* 397-408.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9,* 333-342.

Browne, M. W., MacCallum, R. C., Kim, C. T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods, 7,* 403-421.

Brumbaugh, C. C., & Wood, D. (2013). Mate preferences across life and across the world. *Social Psychological and Personality Science, 4,* 100-107.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: The University of Chicago Press.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98,* 550-558.

Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences, 12,* 1-14.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365-376.

Carnes, N., & Lupu, N. (2016). Do voters dislike working-class candidates? Voter biases and the descriptive underrepresentation of the working class. *American Political Science Review, 110,* 832-844.

Caruso, E. M., Rahnev, D. A., & Banaji, M. R. (2009). Using conjoint analysis to detect discrimination: Revealing covert preferences from overt choices. *Social Cognition, 27,* 128-137.

Cattell R. B. (1966). Psychological theory and scientific method. In R. B. Cattell (Ed.), *Handbook of Multivariate Experimental Psychology* (pp. 1-18). Chicago: Rand McNally.

Cattell, R. B. (1978). *The scientific use of factor analysis*. New York: Plenum.

Center for Empathy in International Affairs. (2016). Empathy in conflict resolution: If, how and when. https://www.centerforempathy.org/wp-content/uploads/2016/06/CEIA-Empathy-in-Conflict-Resolution.pdf

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2019). shiny: Web application framework for R (Version 1.3.2) [Computer software]. https://CRAN.R-project.org/package=shiny

Chow, J. Y., Colagiuri, B., & Livesey, E. J. (2019). Bridging the divide between causal illusions in the laboratory and the real world: The effects of outcome density with a variable continuous outcome. *Cognitive research: Principles and implications. Vol. 4. Cognitive research: Principles and implications* (pp. 1–15).

Christensen, H. T. (1947). Student views on mate selection. *Marriage and Family Living, 9*, 85-88.

Cohen J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34*, 315-346.

Coleman, S. W., Patricelli, G. L., Coyle, B., Siani, J., & Borgia, G. (2007). Female preferences drive the evolution of mimetic accuracy in male sexual displays. *Biology Letters, 3*, 463–466.

Conroy-Beam, D., Goetz, C. D., & Buss, D. M. (2016). What predicts romantic relationship satisfaction and mate retention intensity: Mate preference fulfillment or mate value discrepancies? *Evolution and Human Behavior, 37*, 440-448.

Cottrell, C. A., Neuberg, S. L., & Li, N. P. (2007). What do people desire in others? A sociofunctional perspective on the importance of different valued characteristics. *Journal of Personality and Social Psychology, 92*, 208-231.

Cramer, D. (2003). Facilitativeness, conflict, demand for approval, self-esteem, and satisfaction with romantic relationships. *The Journal of Psychology, 137*, 85-98.

Critcher, C. R., & Zayas, V. (2014). The involuntary excluder effect: Those included by an excluder are seen as exclusive themselves. *Journal of Personality and Social Psychology, 107*, 454-474.

Crowley, B., & Saide, B. (2016). Building empathy in classrooms and schools. *Education Week*. Retrieved from https://www.edweek.org/tm/articles/2016/01/20/building-empathy-in-classrooms-and-schools.html

Davidson, R. J., & Harrington, A. (Eds.). (2002). *Visions of compassion: Western scientists and Tibetan Buddhists examine human nature*. New York: Oxford University Press.

Davis, M. H. (1994). *Empathy: A social psychological approach.* Boulder, CO: Westview Press.

Davis, M. H., & Oathout, H. A. (1987). Maintenance of satisfaction in romantic relationships: Empathy and relational competence. *Journal of Personality and Social Psychology, 53*, 397-410.

de Groot, A. D. (2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. Van der Maas]. *Acta Psychologica, 148*, 188-194.

DeBruine, L. M., Jones, B. C., Little, A. C., Boothroyd, L. G., Perrett, D. I., Penton-Voak, I. S., … Tiddeman, B. P. (2006). Correlated preferences for facial masculinity and ideal or actual partner's masculinity. *Proceedings of the Royal Society of London B: Biological Sciences, 273*, 1355-1360.

Decety, J., & Hodges, S. D. (2004). The social neuroscience of empathy. In P. A. M. Lange (Ed.), *Bridging social psychology*. Hillsdale, NJ: Lawrence Erlbaum.

Decety, J., & Ickes, W. (Eds.). (2009). *The social neuroscience of empathy*. Cambridge, MA: MIT Press.

Delaney, J., Johnson, A. N., Johnson T. D., & Treslan, D. L. (2010). *Students' perceptions of effective teaching in higher education*. St. John's, NL: Memorial University of Newfoundland, Distance Education and Learning Technologies.

Delgado, C., & Guinard, J. X. (2011). How do consumer hedonic ratings for extra virgin olive oil relate to quality ratings by experts and descriptive analysis ratings? *Food Quality and Preference, 22*, 213-225.

DePaulo, B. M., & Friedman, H. S. (1998). Nonverbal communication. In S. T. Fiske, D. Gilbert, & G. Lindzey (Eds.), *The handbook of social psychology* (3rd ed., Vol. 2, pp. 3-40). Boston, MA: McGraw-Hill.

Dolan, C. V., Wicherts, J. M., & Molenaar, P. C. M. (2004). A note on the relationship between the number of indicators and their reliability in detecting regression coefficients in latent regression analysis. *Structural Equation Modeling, 11*, 210-216.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*, 62-68.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*, 52-64.

Eastwick, P. W., & Finkel, E. J. (2008). Sex differences in mate preferences revisited: Do people know what they initially desire in a romantic partner? *Journal of Personality and Social Psychology, 94*, 245-264.

Eastwick, P. W., Finkel, E. J., & Eagly, A. H. (2011). When and why do ideal partner preferences affect the process of initiating and maintaining romantic relationships? *Journal of Personality and Social Psychology, 101*, 1012-1032.

Eastwick, P. W., Joel, S., Molden, D. C., Finkel, E. J., & Carswell, K. L. (2021). Predicting romantic interest during early relationship development: A preregistered investigation using machine learning. *Manuscript under review.*

Eastwick, P. W., Luchies, L. B., Finkel, E. J, & Hunt, L. L. (2014). The predictive validity of ideal partner preferences: A review and meta-analysis. *Psychological Bulletin, 140,* 623-665.

Eastwick, P. W., & Smith, L. K. (2018). Sex-differentiated effects of physical attractiveness on romantic desire: A highly powered, preregistered study in a photograph evaluation context. *Comprehensive Results in Social Psychology, 3*, 1-27.

Eastwick, P. W., Smith, L. K., & Ledgerwood, A. (2019). How do people translate their experiences into abstract attribute preferences? *Journal of Experimental Social Psychology*, *85*, 103837.

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods, 12*, 1-22.

Epskamp, S. (2015). semPlot: Unified visualizations of structural equation models. *Structural Equation Modeling, 22*, 474-483.

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

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Fausset, R. (2017, November 25). I interviewed a white nationalist and fascist. What was I left with? *The New York Times.* Retrieved from https://www.nytimes.com/

Fazio, R. H., Sherman, S. J., & Herr, P. M. (1982). The feature-positive effect in the self-perception process: Does not doing matter as much as doing? *Journal of Personality and Social Psychology, 42*, 404-411.

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual-process Theories in Social Psychology* (p. 97-116). New York: The Guilford Press.

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, 73(3), 421-435.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fiedler, K., Freytag, P., & Meiser, T. (2009). Pseudocontingencies: An integrative account of an intriguing cognitive illusion. *Psychological Review, 116*, 187.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*, 275-297.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Fiske, S. T., & Taylor, S. E. (2008). *Social cognition: From brains to culture*. Boston, MA: McGraw-Hill.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11,* 77-83.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*, 370-378.

Fletcher, G. J. O., Simpson, J. A., & Thomas, G. (2000). The measurement of perceived relationship quality components: A confirmatory factor analytic approach. *Personality and Social Psychology Bulletin, 26*, 340–354.

Fletcher, G. J., Simpson, J. A., Thomas, G., & Giles, L. (1999). Ideals in intimate relationships. *Journal of Personality and Social Psychology*, *76*, 72-89.

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement, 49*, 78-88.

Foldnes, N., & Hagtvet, K. A. (2014). The choice of product indicators in latent variable interaction models: Post hoc analyses. *Psychological Methods*, *19*, 444-457.

Fujita, K. (2011). On conceptualizing self-control as more than the effortful inhibition of impulses. *Personality and Social Psychology Review*, *15*, 352-366.

Gable, S. L., & Reis, H. T. (2010). Good news! Capitalizing on positive events in an interpersonal context. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 42, pp. 195–257). Amsterdam, the Netherlands: Elsevier.

Gable, S. L., Gonzaga, G. C., & Strachman, A. (2006). Will you be there for me when things go right? Supportive responses to positive event disclosures. *Journal of Personality and Social Psychology, 91*, 904-917.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14*, 574-595.

Gerlach, T. M., Arslan, R. C., Schultze, T., Reinhard, S.K., & Penke, L. (2019). Predictive validity and adjustment of ideal partner preferences across the transition into romantic relationships. *Journal of Personality and Social Psychology, 116*, 313-330.

Gilead, M., Trope, Y., & Liberman, N. (2020). Above and beyond the concrete: The diverse representational substrates of the predictive brain. *Behavioral and Brain Sciences, 43, E121*.

Giner-Sorolla, R. (2018). Powering your interaction. Retrieved from https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/

Goldstein, N. J., Vezich, I. S., & Shapiro, J. R. (2014). Perceived perspective taking: When others walk in our shoes. *Journal of Personality and Social Psychology, 106*, 941-960.

Gonzalez, O., & MacKinnon, D. P. (2020). The measurement of the mediator and its influence on statistical mediation conclusions. *Psychological Methods*. Advanced online publication.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*, 148-168.

Goodwin, R., & Tang, D. (1991). Preferences for friends and close relationships partners: A cross-cultural comparison. *The Journal of Social Psychology, 131*, 579-581.

Griffin, D., & Gonzalez, R. (2001). Testing parameters in structural equation modeling: Every "one" matters. *Psychological Methods, 6*, 258-269.

Hainmueller, J. & Hopkins, D. J. (2015). The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants. *American Journal of Political Science, 59*, 529-548.

Hancock, G. R., & Freeman, M. J. (2001). Power and sample size for the root mean square error of approximation test of not close fit in structural equation modeling. *Educational and Psychological Measurement, 61*, 741-758.

Hancock, G. R., & French, B. F. (2013). Power analysis in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 117–159). Charlotte, NC: Information Age.

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin, 137*, 643-659.

Hayes, D. (2005). Candidate qualities through a partisan lens: A theory of trait ownership. *American Journal of Political Science, 49*, 908-923.

Hayes, D. (2010). Trait voting in US senate elections. *American Politics Research, 38*, 1102-1129.

Heider, F. (1958). *The psychology of interpersonal relations.* New York: Wiley.

Heilman, M. E., & Saruwatari, L. R. (1979). When beauty is beastly: The effects of appearance and sex on evaluations of job applicants for managerial and nonmanagerial jobs. *Organizational Behavior & Human Performance, 23*, 360–372.

Hershberger, S. L. (2003). The growth of structural equation modeling: 1994-2001. *Structural Equation Modeling, 10*, 35-46.

Hill, R. (1945). Campus values in mate-selection. *Journal of Home Economics, 37,* 554-558.

Hofmann, W., & Kotabe, H. (2012). A general model of preventive and interventive self-control. *Social and Personality Psychology Compass, 6,* 707-722.

Huang, S. A., Ledgerwood, A., & Eastwick, P. W. (2020). How do ideal friend preferences and interaction context affect friendship formation? Evidence for a domain-general relationship initiation process. *Social Psychological and Personality Science, 11, 226-235*.

Insko, C. A. (1984). Balance theory, the Jordan paradigm, and the Wiest tetrahedron. *Advances in Experimental Social Psychology, 18,* 89-140.

Irving, L. H., & Smith, C. T. (2020). Measure what you are trying to predict: Applying the correspondence principle to the Implicit Association Test. *Journal of Experimental Social Psychology, 86,* 103898.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling, 10,* 128-141.

Jackson, D. L., Gillaspy Jr, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14,* 6-23.

Jenkins, H. M., & Sainsbury, R. S. (1970). Discrimination learning with the distinctive feature on positive or negative trials. In D. Mostofsky (Ed.), *Attention: Contemporary theory and analysis*. New York: Appleton-Century-Crofts, 1970

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79,* 1–17.

Joel, S., Eastwick, P. W., & Finkel, E. J. (2017). Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychological Science, 28,* 1478-1489.

Jones, E. E., & Davis, K. J. (1965). From acts to dispositions: The attribution process in person perception. *Advances in Experimental Social Psychology*, *2*, 220-266.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2018). semTools: Useful tools for structural equation modeling (Version 0.5-1) [Computer software]. https://CRAN.R-project.org/package=semTools

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*, 54-69.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68,* 601-625.

Kaplan, D. (1995). Statistical power in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 100-117). Thousand Oaks, CA, US: Sage Publications, Inc.

Kaplan, M. F. (1973). Stimulus inconsistency and response dispositions in forming judgments of other persons. *Journal of Personality and Social Psychology*, *25,* 58–64.

Kavanagh, L. C., Suhler, C. L., Churchland, P. S., & Winkielman, P. (2011). When it's an error to mirror: The surprising reputational costs of mimicry. *Psychological Science*, *22*, 1274-1276.

Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling*, *12*, 368-390.

Kim, S. S., Kaplowitz, S., & Johnston, M. V. (2004). The effects of physician empathy on patient satisfaction and compliance. *Evaluation & the Health Professions*, 27, 237-251.

Klimecki, O. M. (2019). The role of empathy and compassion in conflict resolution. *Emotion Review*, *11*, 310-325.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press, New York, NY.

Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, *9*, 2-24.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480-498.

Lai, K., & Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods*, *16*, 127-148.

Lampel, A. K., & Anderson, N. H. (1968). Combining visual and verbal information in an impression-formation task. *Journal of Personality and Social Psychology, 9*, 1-6.

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart: The desirability of sociability and competence depends on morality. *Personality and Social Psychology Bulletin*, *42*, 1272-1290.

Lawless, H. T., & Heymann, H. (2010). *Sensory evaluation of food: Principles and practices*. Springer Science & Business Media.

Leary, M. R., & Buttermore, N. R. (2003). The evolution of the human self: Tracing the natural history of self-awareness. *Journal for the Theory of Social Behaviour*, *33*, 365-404.

Ledgerwood, A. (2014). Introduction to the special section on moving toward a cumulative science: Maximizing what our research can tell us. *Perspectives on Psychological Science, 9*, 610–611.

Ledgerwood, A. (2016). Introduction to the special section on improving research practices: Thinking deeply across the research cycle. *Perspectives on Psychological Science, 11*, 661-663.

Ledgerwood, A. (2018). The preregistration revolution needs to distinguish between predictions and analyses. *Proceedings of the National Academy of Sciences, USA, 115*, E10516–E10517.

Ledgerwood, A. (2019). New developments in research methods. In E. J. Finkel & R. F. Baumeister (Eds.), *Advanced Social Psychology* (pp. 39-61). Oxford University Press.

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology, 101*, 1174-1188.

Ledgerwood, A., & Trope, Y. (2010). Attitudes as global and local action guides. In J. Forgas, J. Cooper, & W. Crano (Eds.), *The 12th annual Sydney symposium on social psychology: The psychology of attitude and attitude change* (pp. 39-58). New York: Psychology Press.

Ledgerwood, A., Eastwick, P. W., & Gawronski, B. (2020). Experiences of liking versus ideas about liking. *Behavioral and Brain Sciences*, 43, E136.

Ledgerwood, A., Eastwick, P. W., & Smith, L. K. (2018). Toward an integrative framework for studying human evaluation: Attitudes towards objects and attributes. *Personality and Social Psychology Review, 22*, 378-398.

Lee, Y. J., & Tan, Y. (2013). Effects of different types of free trials and ratings in sampling of consumer software: An empirical study. *Journal of Management Information Systems, 30*, 213-246

Levesque, H. J., Davis, E., & Morgenstern, L. (2011). The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning* (pp. 552–561).

Levine, R. M. (1993, July/August). I feel your pain. *Mother Jones*. Retrieved from https://www.motherjones.com/politics/1993/07/motherjones-ja93-i-feel-your-pain/

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151-173.

Long, E. C., & Andrews, D. W. (1990). Perspective taking as a predictor of marital adjustment. *Journal of Personality and Social Psychology, 59*, 126-131.

Lublin, J. S. (2016, June 21). Companies try a new strategy: Empathy training. *Wall Street Journal*. Retrieved from https://www.wsj.com/

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*, 1122-1135.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201-226.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19-35.

MacCallum, R. C., Browne, M. W. & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

MacCallum, R. C., & Hong, S. (1997). Power analysis in covariance structure modeling using GFI and AGFI. *Multivariate Behavioral Research, 32*, 193-210.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84-99.

Mandel, D. & Lehman, D. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *Journal of Experimental Psychology: General, 127*, 269-285.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70,* 810–819.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181-220.

Marsh, J. K., & Ahn, W. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 334-352.

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology, 6,* 888.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5*, 434-458.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537-563.

McAuliffe, W. H., Carter, E. C., Berhane, J., Snihur, A. C., & McCullough, M. E. (2020). Is empathy the default response to suffering? A meta-analytic evaluation of perspective taking's effect on empathic concern. *Personality and Social Psychology Review, 24*, 141-162.

McNutt, M. (2014). Journals unite for reproducibility. *Science, 346*, 679.

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research, 43*, 1048-1063.

Miller, J., & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science, 11*, 664-691.

Møller, A. (1988). Female choice selects for male sexual tail ornaments in the monogamous swallow. *Nature, 332*, 640–642.

Mooijaart, A. (2003). Estimating the statistical power in small samples by empirical distributions. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 149-156). Tokyo: Springer.

Morelli, S. A., Lieberman, M. D., & Zaki, J. (2015a). The emerging study of positive empathy. *Social and Personality Psychology Compass*, *9*, 57-68.

Morelli, S. A., Sacchet, M. D., & Zaki, J. (2015b). Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. *NeuroImage, 112*, 244-253.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, *89*, 852-863.

Murdoch, C. (2017, February 9). Sapio, the dating app that wants to help smart people hook up. Retrieved from https://mashable.com/2017/02/09/sapio-dating-app/#qWmiR6PmLqqF

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599-620.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Newcomb, T. M. (1953). An approach to the study of communicative acts. *Psychological Review, 60*, 393-404.

Newman, J. P., Wolff, W. T., & Hearst, E. (1980). The feature-positive effect in adult human subjects. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 630-650.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA, 115*, 2600-2606.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615-631.

Nyhan, B. (2015). Increasing the credibility of political science research: A proposal for journal reforms. *PS: Political Science & Politics, 48*, 78-83.

Obama, B. (2006). Obama to graduates: Cultivate empathy. Retrieved from https://www.northwestern.edu/newscenter/stories/2006/06/barack.html

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, USA, 105,* 11087-11092.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867-872.

Osgood, C. E., & Tannenbaum, P. H. (1955). The principle of congruity in the prediction of attitude change. *Psychological Review, 62,* 42-55.

Owen, J. (2018). rhandsontable: Interface to the "Handsontable.js" library (Version 0.3.7) [Computer software]. https://CRAN.R-project.org/package=rhandsontable

Patricelli, G. L., Uy, J. A. C., Walsh, G., & Borgia, G. (2002). Male displays adjusted to female's response. *Nature, 415,* 279–280.

Penev, S., & Raykov, T. (2006). Maximal reliability and power in covariance structure models. *British Journal of Mathematical and Statistical Psychology, 59,* 75-87

Perales, J. C., Catena, A., Shanks, D. R., & González, J. A. (2005). Dissociation between judgments and outcome-expectancy measures in covariation learning: A signal detection theory approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1105.

Pew Research Center Survey (2015, Jan 13). Women and Leadership. Retrieved from http://www.pewsocialtrends.org/2015/01/14/women-and-leadership/st_2015-01-14_ women-leadership-2-01/

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In P. Shaver & M. Mikulincer

(Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). Washington, DC: American Psychological Association.

Preacher, K. J., & Coffman, D. L. (2006, May). Computing power and minimum sample size for RMSEA [Computer software]. Available from http://quantpsy.org/.

Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly*, *26*, 269-281.

Prinz, J. (2011). Is empathy necessary for morality? In A. Coplan, & P. Goldie (Eds.), *Empathy: Philosophical and psychological perspectives* (pp. 211–229). New York: Oxford University Press.

Pronin, E., & Ross, L. (2006). Temporal differences in trait self-ascription: When the self is seen as an other. *Journal of Personality and Social Psychology, 90*, 197-209.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.0) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Atlanta, GA: Sage.

Reinhard, M. A., Greifeneder, R., & Scharmach, M. (2013). Unconscious processes improve lie detection. *Journal of Personality and Social Psychology*, *105*, 721–739.

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331-363.

Rifkin, J. (2009). *The empathic civilization: The race to global consciousness in a world in crisis*. New York: Penguin.

Rosseel, Y. (2012). lavaan: An R Package for structural equation modeling. *Journal of Statistical Software*, *48*, 1-36.

RStudio Team (2018). RStudio: Integrated development for R (Version 1.2.1335) [Computer software]. RStudio, Inc., Boston, MA.

Safire, W. (2008, September 5). Nuance. *The New York Times.* Retrieved from https://www.nytimes.com/

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83-90.

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling, 21*, 149-160.

Savalei, V. (2019). A comparison of several approaches for controlling measurement error in small samples. *Psychological Methods*, *24*, 352-370.

Scarry, E. (1996). The difficulty of imagining other people. In M. C. Nussbaum & J. Cohen (Eds.), *For love of country: Debating the limits of patriotism* (pp. 98-110). Beacon Press.

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, *8*, 379-386.

Schoemann, A. M., Miller, P. M., Pornprasermanit, S., & Wu, W. (2014). Using Monte Carlo simulations to determine power and sample size for planned missing designs. *International Journal of Behavioral Development*, *38*, 471-479.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47,* 609-612.

Shapiro, B. (2017, November). Re "In America's Heartland, the Voice of Hate Next Door" (news article, Nov. 26). *The New York Times*. Retrieved from https://www.nytimes.com/2017/11/27/opinion/nazi-sympathizer.html

Smith, A. (1976). *The theory of moral sentiments* (1st ed.). London, England: Oxford University Press. (Original work published 1759)

Soderberg, C. K., Callahan, S. P., Kochersberger, A. O., Amit, E., & Ledgerwood, A. (2015). The effects of psychological distance on abstraction: Two meta-analyses. *Psychological Bulletin, 141,* 525-548.

Sparks, J., Daly, C., Wilkey, B., Molden, D. C., Finkel, E. J., & Eastwick, P. W. (2020). Negligible evidence that people desire partners who uniquely fit their ideals. Manuscript accepted pending minor revisions at the *Journal of Experimental Social Psychology.*

Spears, N., & Singh, S. N. (2004). Measuring attitude toward the brand and purchase intentions. *Journal of Current Issues & Research in Advertising*, *26*, 53-66.

Spencer-Keyse, J. (2018). Educating empathy: Inspiring students to develop their passions. Retrieved from https://www.brookings.edu/blog/education-plus-development/2018/03/06/educating-empathy-inspiring-students-to-develop-their-passions/

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89*, 845-851.

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences, USA*, *115*, 9210-9215.

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology, 47*, 1249-1254.

Tenold, V. (2018). *Everything you love will burn: Inside the rebirth of white nationalism in America*. Bold Type Books.

Thornhill, R. (1983). Cryptic female choice and its implications in the scorpionfly *Harpobittacus nigriceps*. *The American Naturalist*, *122*, 765-788.

Todd, A. R., & Galinsky, A. D. (2014). Perspective-taking as a strategy for improving intergroup relations: Evidence, mechanisms, and qualifications. *Social and Personality Psychology Compass*, *8*, 374-387.

Todorov, A., Said, C. P., Engel, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*, 455-460.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*, 440-463.

Turban, D. B., & Keon, T. L. (1993). Organizational attractiveness: An interactionist perspective. *Journal of Applied Psychology, 78*, 184-193.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*, 72-81.

Uhlmann, E. L., Zhu, L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition, 126*, 326-334.

Vadillo, M. A., De Houwer, J., De Schryver, M., Ortega-Castro, N., & Matute, H. (2013). Evidence for an illusion of causality when using the Implicit Association Test to measure learning. *Learning and Motivation, 44*, 303-311.

Van der Heijden, H., Verhagen, T., & Creemers, M. (2003). Understanding online purchase intentions: Contributions from technology and trust perspectives. *European Journal of Information Systems, 12*, 41-48.

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98*, 281-300.

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology, 3*, 1. DOI: http://doi.org/10.1525/collabra.74

Vernon, P. (2017). The media today: How not to write about a Nazi. *Columbia Journalism Review.* Retrieved from https://www.cjr.org/the_media_today/new-york-times-nazi-hovater.php

Wakslak, C. J., Nussbaum, S., Liberman, N., & Trope, Y. (2008). Representations of the self in the near and distant future. *Journal of Personality and Social Psychology, 95*, 757-773.

Wall, M. M., & Amemiya, Y. (2001). Generalized appended product indicator procedure for nonlinear structural equation analysis. *Journal of Educational and Behavioral Statistics, 26*, 1-30.

Wallace, B. A., & Shapiro, S. L. (2006). Mental balance and well-being: Building bridges between Buddhism and Western psychology. *American Psychologist, 61*, 690-701.

Wang, Y. A., & Eastwick, P. W. (2020). Solutions to the problems of incremental validity testing in relationship science. *Personal Relationships,* 27, 156–175.

Wang, Y. A., & Todd, A. R. (2020). Evaluations of empathizers depend on the target of empathy. *Journal of Personality and Social Psychology*. Advance online publication. https://doi.org/10.1037/pspi0000341

Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*, 1–17.

Wang, Y. A., Sparks, J., Gonzales, J. E., Hess, Y. D., & Ledgerwood, A. (2017). Using independent covariates in experimental designs: Quantifying the trade-off between power boost and type I error inflation. *Journal of Experimental Social Psychology*, *72*, 118-124.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology/Revue canadienne de psychologie, 19*, 231–241.

Westfall J., & Yarkoni T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS one, 11*, e0152719.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Wickham, H., & Henry, L. (2019). tidyr: Easily tidy data with "spread()" and "gather()" functions (Version 0.8.3) [Computer software]. https://CRAN.R-project.org/package=tidyr

Wojciszke, B., Abele, A. E., & Baryla, W. (2009). Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology, 39*, 973-990.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*, 913-934.

Wood, D., & Brumbaugh, C. C. (2009). Using revealed mate preferences to evaluate market force and differential preference explanations for mate selection. *Journal of Personality and Social Psychology, 96*, 1226–1244.

Yuan, K., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology*, *56*, 93-110.

Yzerbyt, V., Muller, D., Batailler, C., & Judd, C. M. (2018). New recommendations for testing indirect effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology*, *115*, 929-943.

Zaki, J., & Cikara, M. (2015). Addressing empathic failures. *Current Directions in Psychological Science*, *24*, 471-476.

Zaki, J., & Ochsner, K. N. (2012). The neuroscience of empathy: Progress, pitfalls and promise. *Nature Neuroscience*, *15*, 675-680.

Zhang, Y, & Khare, A. (2009). The impact of accessible identities on the evaluation of global versus local products. *Journal of Consumer Research*, 36, 524–537.

Zhang, Z., & Yuan, K.-H. (2018). *Practical statistical power analysis using Webpower and R* (Eds). Granger, IN: ISDSA Press. Retrieved from http://webpower.psychstat.org