Title

Bias in Evaluations in Intergroup and Interpersonal Contexts

Permalink

https://escholarship.org/uc/item/5gh5t2v0

Author

da Silva Frost, Aline

Publication Date

2024

Peer reviewed|Thesis/dissertation

Bias in Evaluations in Intergroup and Interpersonal Contexts

By

ALINE DA SILVA FROST
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 Eastwick

Andrew Todd

Committee in Charge

2024

Acknowledgements

First, I would like to thank my sister, Dr. Paula da Silva Frost, my parents, Juraciara Pereira da Silva and Marcos da Silva Frost, and my whole family and community. I am extremely grateful to my amazing advisor and committee members, Dr. Alison Ledgerwood, Dr. Paul Eastwick, and Dr. Andrew Todd, for all their support, advice and invaluable feedback. Finally, I would like to thank my peers, Dr. Samuel Klein, Dr. Christopher Coleman, Dr. Andre Wang and the other members of ASC Lab, along with all my collaborators. Words cannot express my gratitude.

O correr da vida embrulha tudo,
a vida é assim: esquenta e esfria,
aperta e daí afrouxa,
sossega e depois desinquieta.
O que ela quer da gente é coragem

João Guimarães Rosa

Table of Contents

Abstract

Biases in evaluations are an essential part of human experience, and unsurprisingly, they have been studied across multiple and often disparate literatures. The three papers I present here further our understanding of biases by integrating across different theoretical frameworks and research areas. In Chapter 1, I present advances in research methods and practices, which include discussions on statistical power, how to measure and control Type I error, and preregistrations. This paper sets the basis for the how and why of the methods used in the subsequent chapters. In Chapter 2, I dive into intergroup contexts to investigate a frequently overlooked confound in the literature of implicit bias: the exemplar-category confound. In common research practices in the field, implicit measures include presentations of specific exemplars (e.g., faces of Black people), whereas explicit measures often focus on responses to abstract social categories (e.g., feeling thermometer towards the category Black people). Results of four experiments suggest that previously obtained implicit-explicit dissociations using the Implicit Association Test may be at least partly driven by the exemplar-category confound. In Chapter 3, I turn into interpersonal contexts to investigate evaluative biases in the romantic relationships literature. First, I successfully developed a paradigm to reliably manipulate people's evaluations of traits, and afterwards, I used this paradigm to conduct the first experimental tests of four key theoretical accounts. In conclusion, this work develops and disseminates new methods and paradigms, challenges longstanding assumptions in the field, and furthers our understanding of biases and their consequences.

Chapter 1

A Tutorial on Improving Research Methods and Practices

Abstract

This article provides an accessible tutorial with concrete guidance for how to start improving research methods and practices in your lab. Following recent calls to improve research methods and practices within and beyond the borders of psychological science, resources have proliferated across book chapters, journal articles, and online media. Many researchers are interested in learning more about cutting-edge methods and practices, but are unsure where to begin. In this tutorial, we describe specific tools that help researchers calibrate their confidence in a given set of findings. In Part I, we describe strategies for assessing the likely statistical power of a study, including when and how to conduct different types of power calculations, how to estimate effect sizes, and how to think about power for detecting interactions. In Part II, we provide strategies for assessing the likely Type I error rate of a study, including distinguishing clearly between data-independent ("confirmatory") and data-dependent ("exploratory") analyses and thinking carefully about different forms and functions of preregistration.

Keywords: statistical power, pre-registration, pre-analysis plan, open science, replicability, positive predictive value

A tutorial on improving methods and practices

In recent years, psychology has been the forefront of a broad movement across scientific disciplines to improve the quality and rigor of research methods and practices (Begley & Ellis, 2012; Button et al., 2013; Ledgerwood, 2016; McNutt, 2014; Nosek, Spies, & Motyl, 2012; Nyhan, 2015; see Spellman, 2015, for a helpful synopsis). The field as a whole is changing: Conversations about improving research practices have become mainstream, journals and societies are adopting new standards, and resources for improving methods and practices have proliferated across journal articles, book chapters, and online resources like blogs and social media (Simons, 2018). As attention to methodological issues has surged, researchers have become increasingly interested in understanding and implementing methodological tools that can maximize the knowledge they get from the work that they do.

At the same time, for the average researcher, it can be daunting to approach this new wealth of resources for the first time. You know that you want to understand the contours of recent developments and to learn as much as possible from the research you do, but where do you even begin? We think that one of the most important methodological skills to develop is how to calibrate your confidence in a finding to the actual strength of that finding.

In this tutorial, we seek to provide a toolbox of strategies that can help you do just that. If a finding is strong, you want to have a relatively high level of confidence in it. In contrast, if a finding is weak, you want to be more skeptical or tentative in your conclusions. Having too much confidence in a finding can lead you to waste resources chasing and trying to build on an effect that turns out to have been a false positive, thereby missing opportunities to discover other true effects. Likewise, having too little confidence in a finding can lead you to miss opportunities to build on solid and potentially important effects. Thus, in order to maximize what we learn from the work that we do as scientists, we want to have a good sense of how much we learn from a given finding.

We divide this tutorial into two main parts. The first part will focus on how to estimate statistical power, which refers to the likelihood that a statistical test will correctly detect a true effect if it exists (i.e., the likelihood that if you are testing a real effect, your test statistic will be significant). The second part will focus on Type I error, which refers to the likelihood that a statistical test will incorrectly detect a null effect (i.e., the likelihood that if you are testing a null effect, your test statistic will be significant). Arguably, one of the central problems giving rise to the field's so-called "replicability crisis" is that researchers have not been especially skilled at assessing either the statistical power or the Type I error rate of a given study—leading them to be overly confident in the evidential value and replicability of significant results (see Anderson, Kelley, & Maxwell, 2017; Lakens & Evers, 2014; Ledgerwood, 2018; Nosek, Ebersole, DeHaven, & Mellor, 2018a; Open Science Collaboration, 2015; Spellman, Gilbert, & Corker, 2017; Pashler & Wagenmakers, 2012). For example, Bakker et al. (2012) estimated the average statistical power in psychological experiments to be only 35%, and even large studies may have lower statistical power than researchers intuitively expect when measures are not highly reliable (see Kanyongo et al., 2007; Wang & Rhemtulla, in press). Meanwhile, common research practices can inflate the Type I error rate of a statistical test far above the nominal alpha (typically p < .05) selected by a researcher (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011; Wang & Eastwick, in press).

Importantly, you need a good estimate of both quantities—statistical power and Type I error—in order to successfully calibrate your confidence in a given finding. That's because both quantities influence the *positive predictive value* of a finding, or how likely it is that a significant result reflects a true effect in the population.[1]

For example, imagine that in the course of a typical year, a researcher has ten ideas that happen to be correct and ten ideas that happen to be incorrect (that is, she tests ten effects that are in fact true effects in the population and ten that are not). Let's focus on what happens to the correct ideas first. As illustrated in Figure 1, the statistical power of the researcher's studies determines how many of these true effects will be detected as significant. If the studies are powered at 40% (left side of Figure 1), four out of ten studies will correctly detect a significant effect, and six out of ten studies will fail to detect the effect that is in fact present in the population. If the studies are powered at 90% (right side of Figure 1), nine out of ten studies will correctly detect the significant effect, and only one will miss it and be placed in the file drawer.

However, statistical power is only part of the story. Not all ideas are correct, and so let's focus now on the ten ideas that happen to be incorrect (that is, she tests ten effects that are in fact null effects in the population). As illustrated in Figure 1, the Type I error rate of the researcher's studies determines how many of these null effects will be erroneously detected. If the studies have a Type I error rate of 30%, three of the ten null effects will be erroneously detected as significant, and the other seven will be correctly identified as non-significant.[2]

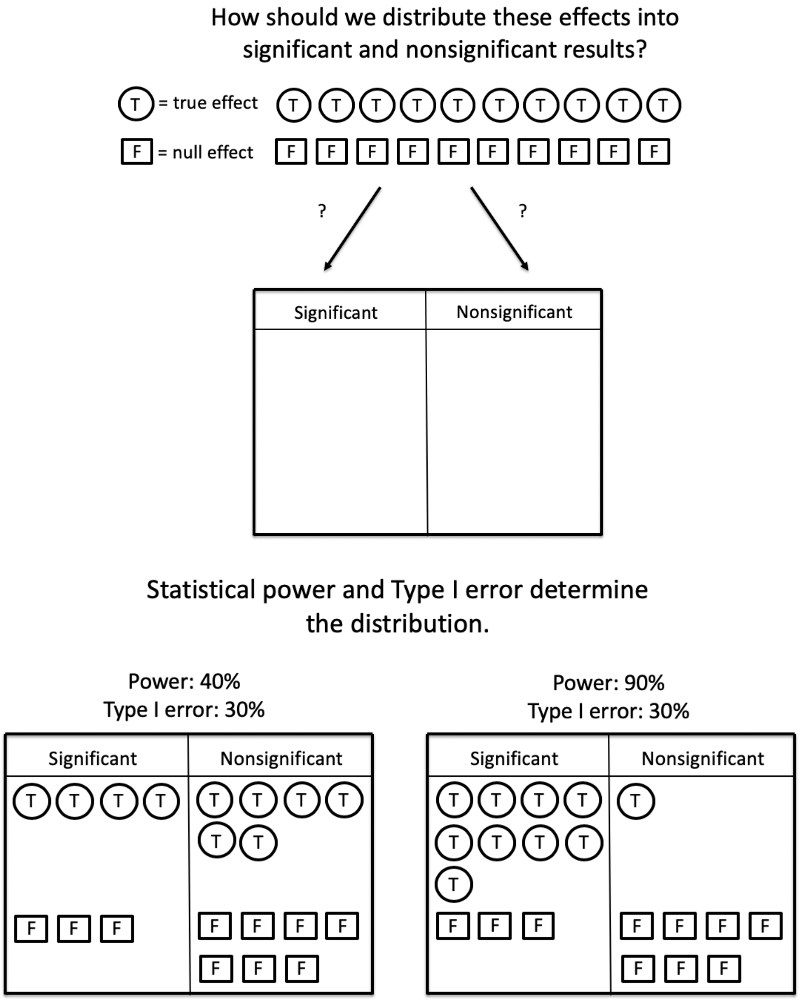


Figure 1. Consider the case of a researcher testing 10 true effects and 10 false effects. Perhaps they will follow up or publish significant results but leave nonsignificant results in a file drawer. The statistical power and Type I error rate of the studies will determine how the effects are sorted into a set of significant results (follow up!) and a set of nonsignificant results (file drawer). Notice that because power is higher in the scenario on the right (vs. left), the likelihood that any one of the significant findings reflects a true effect in the population is also higher.

The researcher, of course, does not know whether the effects are real or null in the population; she only sees the results of her statistical tests. Thus, what she really cares about is how likely she is to be right when she reaches into her pile of significant results and declares: "This is a real effect!" In other words, if she publishes or devotes resources to following up on one of her significant effects, how likely is it to be a correctly detected true effect, rather than a false positive? Notice that the answer to this question about the positive predictive value of a study depends on both statistical power and Type I error rate. In Figure 1, the positive predictive value of a study is relatively low when power is low (on the left side): The likelihood that a significant result in this pool of significant results reflects a true effect is 4 out of 7, or 57%. In contrast, the positive predictive value of a study is higher when power is higher (on the right side of Figure 1): The likelihood that a significant result in this pool of significant results reflects a true effect is 9 out of 12, or 75%. Thus, if we are to understand how much to trust a significant result, we want to be able to gauge *both* the likely statistical power and the likely Type I error rate of the study in question.

At this point, readers may wonder about the tradeoff between statistical power and Type I error, given that the two are related. For example, one way to increase power is to set a higher alpha threshold for significance testing (e.g., $p < .10$ instead of $p < .05$), but this practice will also increase Type I error. However, there are other possible ways to increase power that do not affect the Type I error rate (e.g., increasing sample size, improving the reliability of a measure)—and it is these strategies that we discuss below. More broadly, in this tutorial, we focus on providing tools to assess (1) the likely statistical power and (2) the likely Type I error rate of a study result, and offer guidance for how to increase statistical power or constrain Type I error for researchers who want to be able to have more confidence in a given result.

Part I: How to Assess Statistical Power

Develop Good Intuitions about Effect Sizes and Sample Sizes

One simple but useful tool for gauging the likely statistical power of a study is a well-developed sense of the approximate sample size required to detect various effects. Think of this as building your own internal power calculator that provides rough, approximate estimations.

You want to be able to glance at a study and think to yourself: “Hmm, that’s a very small sample size for studying this type of effect with this sort of design—I will be cautious about placing too much confidence in this significant result” or “This study is likely to be very highly powered—I will be relatively confident in this significant result.” In other words, it’s useful to develop your intuitions for assessing whether you’re more likely to be in a world that looks more like the left side of Figure 1 or in a world that looks more like the right side.

How do you build this internal calculator? You can start by memorizing some simple benchmarks. For a simple, two-condition, between-subjects study, the sample size required to detect a medium effect size of *d* = .50 with 80% power is about N = 130 (65 participants per condition; see Figure 2. Notice that *d* = .50 is equivalent to *r* = .24 and *h2* = .06). To detect a large effect of *d* = .80 with 80% power requires about N = 50 (25 per condition). And to detect a small effect size of *d* = .20 with 80% requires about N = 800 (400 per condition; Faul, Erdfelder, Lang & Buchner, 2007).

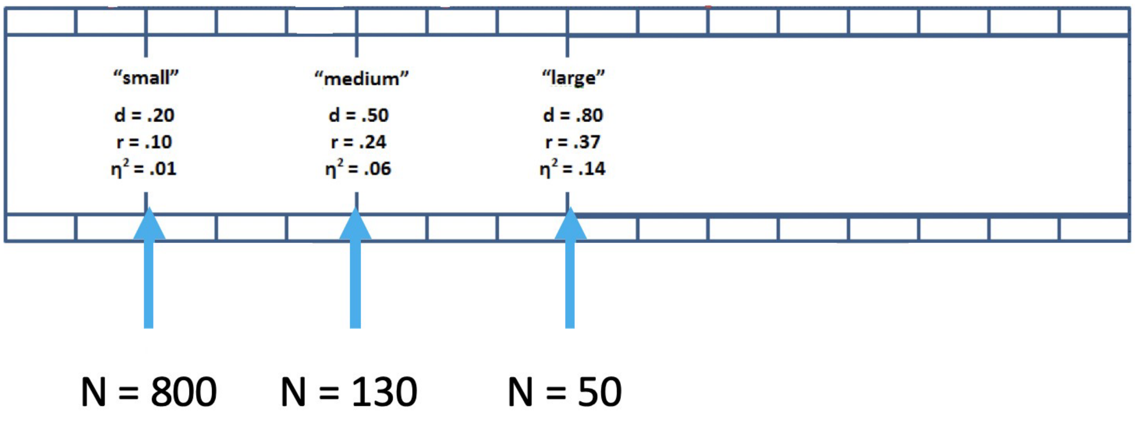


Figure 2. Sample sizes needed to achieve 80% power in a two-condition, between-subjects study. This figure helps you organize visually the effect size intuitions.

Next, start developing your sense of how big such effects really are. Cohen (1992) set a medium effect size at *d* = .50 to "represent an effect likely to be visible to the naked eye of a careful observer" (p. 156), so a medium-sized effect is one that we might observe simply by watching people closely. A large effect size of *d* = .80 is typically an effect that even a casual observer would notice (for example, the correlation between relationships satisfaction and breakup is approximately this magnitude; Le et. al., 2010). And a small effect size of *d* = .20 is typically too small to be seen with the naked eye (for example, if you're interested in testing a counterintuitive prediction that would be surprising to most people, it is likely to be a small effect if it is true). Pay attention to effect size estimates from meta-analyses and very large studies in your area of research to hone your intuitions within your particular research area.

Finally, keep in mind that interactions can require much larger sample sizes, depending on their shape. For example, imagine that you are interested in powering a two-group study to detect a medium-sized effect. You conduct Study 1 with a total sample size of N = 130 and find that indeed, your manipulation (let's call it Factor A) significantly influences your dependent measure. Next, you want to know if Factor B moderates this effect. How many participants do you need to have 80% power to detect an interaction in Study 2, where you manipulate both Factor A and Factor B in a 2 x 2 design?

As Table 1 illustrates, the answer depends on the shape of the interaction you expect (see Giner-Sorolla, 2018; Ledgerwood, 2019). If you expect a knockout interaction (i.e., you think the effect you saw in Study 1 will appear in one condition of Factor B and disappear in the other), it turns out you need to *quadruple* the sample size you had in Study 1 to have 80% power to detect the interaction in Study 2. If you expect a perfect cross-over interaction (i.e., you think the effect you saw in Study 1 will appear in one condition of Factor B and reverse completely in the other), you need the *same* sample size you had in Study 1 to have 80% power to detect the interaction in Study 2 (although you will probably want to double the sample size to provide 80% power to detect each of the simple main effects). And if you expect a 50% attenuation (i.e., you think the effect you saw in Study 1 will appear in one condition of Factor B and be reduced to half its original size in the other), you need about *14 times* the sample size you used in Study 1.

Table 1.

Rules of thumb for powering a 2x2 between-subjects Study 2 that seeks to moderate a main effect observed in Study 1.

Expected shape of interaction	**Required total sample size for Study 2**	**Example: N needed if Study 1 tested an effect of *d* = .50**
Cross-over	2x N in Study 1[a]	N ≈ 260
Knockout	4x N in Study 1	N ≈ 520
50% attenuation	14x N in Study 1	N ≈ 1820

Note. In this illustration, Study 1 is powered to provide 80% power to detect a main effect. In Study 2, a researcher wants to test whether the effect observed in Study 1 is moderated by a second variable in a 2x2 between-subjects factorial design.

Strategies For a Planned Study

Conduct an A Priori Power Analysis

The rules of thumb described above can be useful, but when you are planning your own study, you can conduct a formal, *a priori* power analysis to decide how many participants you need in order to achieve your desired level of power. In an *a priori* power analysis, you input your desired level of power (e.g., 80%), your planned statistical test (e.g., a *t*-test comparing two between-subjects conditions), and your estimated effect size (e.g., *d* = .40), and the program tells you the necessary sample size (*e.g.*, N = 200). The central challenge in this kind of power analysis is to identify a good estimate of the expected effect size.

Getting a good estimate of the expected effect size can be tricky for multiple reasons.

First, effect size estimates (like any estimate) will fluctuate from one study to the next, especially when sample sizes are smaller (see Ledgerwood, Soderberg, & Sparks, 2017, Figure 1, for an illustration). In other words, an effect size estimate from any given study can underestimate or overestimate the true size of an effect. Second, publication bias tends to inflate the effect sizes reported in a given literature. Historically, significant results were (and continue to be) more likely to be published, and null effects are more likely to be shuttled to a file drawer rather than shared with the scientific community (Anderson, Kelley, & Maxwell, 2017; Ledgerwood, 2019; Rosenthal, 1979). Because any given study can underestimate or overestimate an effect size, and because overestimates are more likely to hit significance, publication bias effectively erases a sizable portion of underestimates from the literature. Therefore, a published effect size estimate is more likely to be an overestimate than an underestimate. And, when one averages the effect size estimates that do make it into the published literature, that average is usually too high (Anderson et al., 2017).

To get around these issues, we have two options: a large study or a meta-analysis. In both cases, it is important to consider publication bias. The first option is to find an estimate of a similar effect from a large study (e.g., a total sample size of approximately N = 250 or larger for estimating a correlation between two variables or a mean difference between two groups; Schönbrodt & Perugini, 2013). If you find such a study, ask yourself whether the paper would have been published if it had different results (e.g., a paper that describes its goal as estimating the *size* of an effect may be less affected by publication bias than a paper that describes its goal as demonstrating the *existence* of an effect). The second option is to find an estimate of a similar effect from a meta-analysis. Meta-analyses aggregate results of multiple studies, so their estimates ought to be more accurate than an estimate from a single study. However, because they also sample studies from the literature, they can overestimate the size of an effect. For this reason, look for meta-analysis that carefully model publication bias (e.g., by using sensitivity analyses that employ various models of publication bias to produce a range of possible effect size estimates; McShane, Bockenholt, & Hansen, 2016; see Ledgerwood, 2019, for a fuller discussion).

By identifying a good estimate of the expected effect size, such as those from large studies and meta-analyses that account for publication bias, we can conduct a priori power analyses that will be reasonably accurate. You can conduct an *a priori* power calculation using many simple-to-use programs, such as G*Power (Faul, Erdfelder, Lang & Buchner, 2007); PANGEA (Westfall, 2016) for general ANOVA designs; and pwrSEM (Wang & Rhemtulla, in press) for structural equation models.

Of course, sometimes it is not possible to identify a good effect size estimate from a large study or meta-analysis that accounts for publication bias. If your only effect size estimate is likely to be inaccurate and/or biased (e.g., an effect size estimate from a smaller study), you can use a program that accounts for uncertainty and bias in effect size estimates (available online at

https://designingexperiments.com/shiny-r-web-apps under the penultimate heading, "Bias and Uncertainty Corrected Sample Size for Power" or as an R package; see Anderson et al., 2017).

Identify the Biggest Sample Size Worth Collecting

In other situations, you are simply not sure what effect size to expect—perhaps you are starting a brand-new line of research, or perhaps the previous studies in the literature are simply too small to provide useful information about effect sizes. A useful option in such cases is to identify the smallest effect size of interest (often abbreviated SESOI) and use that effect size in your power calculations (Lakens, 2014). In other words, if you only care about the effect if it is at least medium in size, you can power your study to detect an effect of $d = .50$.

Basic researchers often feel reluctant to identify a SESOI, because they often care about the direction of an effect regardless of its size (e.g., competing theories might predict that a given manipulation will increase or decrease levels on a given dependent measure, and a basic researcher might be interested in either result regardless of the effect size). However, with a minor tweak, the basic concept becomes useful to everyone. If identifying a SESOI feels difficult, identify instead the largest sample you would be willing to collect to study this effect.

Let's call this the "Biggest Sample Size Worth Collecting," or BSSWC. For example, if you decide that a given research question is worth the resources it would take to conduct a two-group experiment with a total of N = 100 participants, you are effectively deciding that you are only interested in the effect if it is at least $d = .56$ (the effect size that N = 100 would provide about 80% power to detect). Notice, then, that a BSSWC of 100 participants is equivalent to a SESOI of $d = .56$—they are simply two different ways of thinking about the same basic idea. Notice, too, that it is worth making the connection between these two concepts explicit. For example, a social psychologist might consider whether the effect they are interested in studying is larger or smaller than the average effect studied in social-personality psychology ($r = .21$ or $d = .43$; see Richard, Bond, & Stokes-Zoota, 2003). Unless they have a reason to suspect their effect is much larger than average, they may not want to study it with a sample size of only N = 100 (because they will be under-powered; see Figure 1).

Conserve Resources When Possible: Sequential Analyses

Once you have determined the maximum sample size you are willing to collect (either through an *a priori* power analysis or by determining the maximum resources you are willing to spend on a given study), you can conserve resources by using a technique called *sequential analysis*. Sequential analyses allow you to select *a priori* the largest sample size you are willing to collect if necessary as well as middle points where you would stop data collection earlier if you could (Lakens & Evers, 2014; Proschan, Lan, & Wittes, 2006). For example, if you have a wide range of plausible effect size estimates and are unsure about how many participants to run, but you know you are willing to collect a total sample size of N = 600 to detect this particular effect, a sequential analysis may be the best option. Sequential analyses are planned ahead of time, before looking at your results, and preserve a maximum Type I error rate of 5%. In contrast, if you check your data multiple times *without* using a formal sequential analysis, Type I error rates inflate (see Sagarin, Ambler, & Lee, 2014; Simmons et al., 2011).

To conduct a sequential analysis, you would first decide how many times you will want to check your data before reaching your final sample size—in our example, N = 600. Each additional check will reduce power by a small amount in exchange for the possibility of stopping early and conserving resources. Imagine that you decide to divide your planned sample into three equal parts, so that you conduct your analysis at n = 200, n = 400 and N = 600 participants (you can follow this example in Table 2). You would then pause data collection at each of these points and check the results. If the *p*-value of the analysis is less than a predetermined alpha threshold (see Table 2), you would determine that the test is statistically significant. If it is greater than the threshold, you would continue collecting data up until the final sample size of N = 600.

For instance, imagine that you collect the first planned set of 200 participants, or 33% of the total N. You pause data collection and analyze the data; if the p-value of the focal analysis is below the first alpha threshold of .017, the result is significant and you can stop data collection early (saving 400 participants). If the p-value of the focal analysis is not below .017, you continue to collect data from 200 more participants. When you check the results again, if the p-value is below the second alphas threshold of .022, the result is significant and you can stop data collection early (saving 200 participants). If the p-value is not below .022, you continue to collect data from 200 more participants and then conduct the focal analysis one final time on the total sample of 600 participants. If the p-value is below .028, the result is significant. If it is not below .028, the result is not significant. In either case, you stop collecting data because you have reached your final planned sample size.

Importantly, by computing specific, adjusted alpha thresholds depending on the planned number of stopping points, sequential analysis enables researchers to check the data multiple times during data collection while holding the final Type I error rate at a maximum of .05. This allows you to balance the goals of maximizing power and conserving resources. Table 2 provides the alpha thresholds for common sequential analyses where a researcher wants to divide their total planned sample (of any size) into 2-4 equal parts and hold their Type I error rate at .05 or below (for an example of how to write up a sequential analysis, see Sparks & Ledgerwood, 2017). If you want to stop more frequently or at unevenly spaced points, you can use the GroupSeq R package and step-by-step guide provided by Lakens (2014; resources available at https://osf.io/qtufw/) to compute other alpha thresholds.

Table 2.

Alpha Thresholds for Sequential Analyses

Divide your sample size into _ equal parts	Stop at		Alpha Threshold	Decision Guide	
	percent of total N	example (total N=600)			
2	50%	300	.025	p<.025?	if yes, significant. if no, continue collection
	100%	600	.034	p<.034?	if yes, significant if no, it is not significant
3	33%	200	.017	p<.017?	if yes, significant. if no, continue collection
	66%	400	.022	p<.022?	if yes, significant if no, continue collection
	100%	600	.028	p<.028?	if yes, significant if no, it is not significant
4	25%	150	.013	p<.013?	if yes, significant if no, continue collection
	50%	300	.016	p<.016?	if yes, significant if no, continue collection
	75%	450	.020	p<.020?	if yes, significant if no, continue collection
	100%	600	.025	p<.025?	if yes, significant if no, it is not significant

Note. Once you have planned your total sample size and how many times you will want to stop and check the results, use this table to determine the alpha cut-off thresholds you will use to determine significance at each planned analysis time point.

Consider Multiple Approaches to Boosting Power

After conducting an *a priori* power analysis, you may find that to have your desired level of power, you need a larger sample size than you initially imagined. However, it is not always possible or practical to collect large sample sizes. You may be limited by the number of participants available to you (especially when studying hard-to-reach populations) or by the finite money and personnel hours that you have to spend on collecting data. Whatever the situation, all researchers face trade-offs and constraints based on resources.

Given such constraints, it is often useful to consider multiple approaches to boosting the power of a planned study. When possible and appropriate, making a manipulation within-subjects instead of between-subjects can dramatically boost the power of an experiment (see

Greenwald, 1976; Rivers & Sherman, 2018). Likewise, you can increase power by strengthening a manipulation and by improving the reliability of your measures (see e.g., Ledgerwood & Shrout, 2011). In addition, it is sometimes possible to select ahead of time a planned covariate that correlates strongly with the dependent measure of interest (e.g., measuring extraversion as a covariate for a study that examines self-esteem as the focal dependent variable; see Wang, Sparks, Gonzales, Hess, & Ledgerwood, 2017). Finally, one of the most exciting developments in the "cooperative revolution" created by the open science movement is the proliferation of opportunities for large-scale collaborations (e.g., the Psychological Science Accelerator, ManyLabs, ManyBabies, and StudySwap; see Chartier, Kline, McCarthy, Nuijten, Dunleavy, & Ledgerwood, 2018). When it simply is not feasible to study the research question you want to study with high statistical power, consider collaborating across multiple labs and aggregating the results.

Strategies for an Existing Study

Conduct a Sensitivity Analysis

When you want to assess the statistical power of an existing study (e.g., a study published in the literature or a dataset you have already collected), you can conduct a type of power analysis called a *sensitivity analysis* (Cohen, 1998; Erdfelder, Faul & Buchner, 2005). In a sensitivity analysis, you input the actual sample size used in the study of interest (e.g., N = 60), the statistical test (e.g., a *t*-test comparing two between-subjects conditions), and a given level of power (e.g., 80%), and the program tells you the effect size the study could detect with this level of power (e.g., *d* = .74). The central goal for this kind of power analysis is to provide a good sense of the range of effect sizes that an existing study was adequately powered to detect.

For example, perhaps you have already conducted a study in which you simply collected as many participants as resources permitted, and you ended up with a total sample of N = 164 participants in a two-group experimental design. You could conduct sensitivity analyses to determine that your study had 60% power to detect an effect size of *d* = .35 and 90% power to detect an effect of *d* = .51. Armed with the effect size intuitions we discussed in an earlier section, you could then ask yourself whether the effect size you are studying is likely to be on the smaller side or on the larger side (e.g., is it an effect that a careful observer could detect with the naked eye?). By thinking carefully about this information, you can gauge the likely statistical power of your study (e.g., is it more like the left side or the right side of Figure 1?) and calibrate your confidence in the statistical result accordingly.

Don't Calculate "Post-Hoc" or "Observed" Power

Many types of power analysis software also provide an option for computing power called *post hoc power* or *observed power*. This type of power analysis is highly misleading and should be avoided (Gelman & Carlin, 2014; Hoenig & Heisey, 2001). In a post hoc power analysis, you input the effect size estimate from a study as if it is the true effect size in the population. However, as discussed earlier, a single study provides only one estimate of the true population effect size, and this estimate tends to be highly noisy: It can easily be far too high or far too low (see Figure 1 in Ledgerwood, Soderberg, & Sparks, 2017; Schönbrodt & Perugini, 2013). Furthermore, because researchers tend to be more interested in following up on and publishing significant results, and because a study is more likely to hit significance when it overestimates (vs. underestimates) an effect size, researchers are especially likely to conduct post hoc power analyses with overestimated effect size estimates.

The result of using post-hoc power is an illusion of a precise power estimate that in fact is (1) highly imprecise and (2) redundant with the *p*-value of the study in question. In other words, post-hoc power or observed power appears to provide a new piece of very precise information about a study, when in fact it provides an already known piece of imprecise information. It is loosely akin to attempting to gauge the likelihood that a coin flip will result in "heads" rather than "tails" based on flipping the coin, observing that the result is "heads," and then deciding based on this observation that the coin flip must have been very likely to result in the outcome you saw. Thus, post-hoc power or observed power ultimately *worsens* your ability to calibrate your confidence to the strength of a result.

Part II: How to Assess Type I Error Rates

As Figure 1 illustrates, if we want to correctly calibrate our confidence in a significant result, we want to be able to gauge not only the likely statistical power of the test in question, but also its likely Type I error rate. Researchers often assume that their Type I error rate is simply set by the alpha cut-off against which a *p*-value is compared (traditionally, $p < .05$). In reality, however, the likelihood of mistakenly detecting a significant effect when none exists in the population can be inflated beyond the nominal alpha rate (.05) by a number of factors.

Understand How Data-Dependent Decisions Inflate the Type I Error Rate

Perhaps most importantly, the Type I error rate can inflate—often by an unknown amount—when the various decisions that a researcher makes about how to construct their dataset and analyze their results are informed in some way by the data themselves. Such decisions are called *data-dependent* (or often "exploratory," although this term can have multiple meanings and so we avoid it here for the sake of clarity). For example, if a researcher decides whether or not to continue collecting data based on whether their primary analysis hits significance, the Type I error rate of that test will inflate a little (if they engage in multiple rounds of such "optional stopping," Type I error can increase substantially; see Sagarin, Ambler, & Lee, 2014). Likewise, running an analysis with or without a variety of possible covariates until one hits significance can inflate Type I error (see Wang et al., 2017), as can testing an effect with three slightly different dependent measures and reporting only the ones that hit significance. In fact, even the common practice of conducting a 2 x 2 factorial ANOVA and reporting all effects (two main effects and an interaction) has an associated Type I error rate of about 14% rather than the 5% researchers typically assume (see Cramer et al., 2016). In all of these cases, the problem arises because there are multiple possible tests that a researcher could or does run to test their research question (e.g., a test on a subsample of 100 and a test on a subsample of 200; a test on one dependent measure versus a test on a different dependent measure). When the decision about which test to run and report is informed by knowledge of the dataset in question, the Type I error rate starts to inflate (see Gelman & Loken, 2014; Simmons et al., 2011).

Clearly Distinguish between Data-Dependent and Data-Independent Analyses with a Pre- Analysis Plan

Of course, the fact that data-dependent analyses inflate Type I error does not mean that you should never let knowledge of your data guide your decisions about how to analyze your data. Data-dependent analyses are important to get to know your data and to help generate new hypotheses and theories. Moreover, in some research areas, it is difficult or impossible to analyze a dataset without already knowing something about the data (e.g., political science studies of election outcomes; Gelman & Loken, 2014). Data-dependent analyses are often extremely useful, but we want to *know* that an analysis is data-dependent so that we can calibrate our confidence in the result accordingly.

Thus, an important tool to have in your toolkit is the ability to distinguish clearly between data-dependent and data-independent analyses. A well-crafted *pre-analysis plan* allows you to do just that. A pre-analysis plan involves selecting and writing down ahead of time the various researcher decisions that will need to be made about how to construct and analyze a dataset, before looking at the data. Writing down the decisions ahead of time is important to circumvent human biases in thinking and memory (Chaiken & Ledgerwood, 2011; Nosek, Spies, & Motyl, 2012)—after looking at the data, it is very easy to convince ourselves we actually intended to do these particular tests and make these particular decisions all along. By creating a record of which decisions were in fact data-independent, pre-analysis plans allow researchers to distinguish between data-dependent and data-independent analyses. For example, if you write down ahead of time that you will include a carefully chosen covariate in your analysis and you follow that plan, you can rest assured that you have not unintentionally inflated your Type I error rate (Wang et al., 2017). On the other hand, if you decide after looking at the data to include a different covariate or none at all, you can calibrate your confidence in that data-dependent analysis accordingly (e.g., being more tentative about that result until someone can test if it replicates). Pre-analysis plans thus enable you to plan your analyses with a constrained Type I error rate, allowing you to know what this rate is. However, the plan is not a guarantee for keeping an alpha level below the desired rate (usually .05). If you plan multiple comparisons (e.g., you plan to test all effects in a 2-way ANOVA; Cramer et al., 2016) or inappropriate statistical tests (e.g., you use multiple regression rather than latent variables to test the incremental validity of a psychological variable, which can produce spurious results due to measurement error; see Westfall & Yarkoni, 2016; Wang & Eastwick, in press), your Type I error rate may be higher than you imagine. Also, it is important not to follow the plan blindly, and always check if the assumptions of a statistical test are met given the data.

When constructing a pre-analysis plan for the first time, it is often useful to start with a template designed for your type of research. For example, psychological researchers conducting experiments often find the AsPredicted.org template useful because it clearly identifies the most common researcher decisions that an experimental psychologist will need to make and provides clear examples of how much detail to include about each one. Other templates are available on OSF (see https://osf.io/zab38/), or you can create your own tailored to your own particular research context (see Table 3 for an example). Don't be surprised to find that you forget to record some researcher decisions the first time you create a pre-analysis plan for a given line of research. It can be hard to anticipate all the decisions ahead of time. But even when a pre-analysis plan is incomplete, it can help you clearly identify those analyses that were planned ahead of time and those that were informed in some way by the data.

Table 3.

Common Decisions to Specify in a Pre-Analysis Plan

Consider Describing:	**Example:**
Planned sample size and stopping rule	Target total N = 200 We will collect data until Qualtrics indicates that there are completed surveys from 200 participants.
Inclusion criteria	University students 18 and older who have not participated in a previous study in this line of work will be allowed to participate.
Exclusion criteria	Participants will be excluded if they respond incorrectly to the attention check at the end of the study, as coded by a researcher blind to the rest of the data. UPDATED 10/20/2019 after opening the data file but before running any analyses: We noticed that 2 participants spent less than 2 seconds reading the screen that displayed the manipulation, whereas everyone else spent at least 30 seconds, so we decided to exclude these 2 participants.
Manipulation(s) and conditions	Consensus information (2 between-subjects conditions): Participants read that 70% of students at their university support vs. oppose a new bike law.

Predictor(s) and how they will be constructed	N/A
Dependent measure(s) and how they will be constructed	Primary/Focal DV: Participants' own attitudes toward the bike law (average of the five-item scale) Additional DV: Attitude strength (average of the four-item scale)
Any planned covariates	N/A
Planned statistical tests involving specific operational variables	Primary/Focal analysis: Independent t-test (two-tailed) examining the effect of condition on participants' attitudes toward the bike law. Additional analysis: Independent t-test (two-tailed) examining the effect of condition on participants' attitude strength.
Any planned follow-up or subgroup analyses	No
Any plan for Type I error control (e.g., for multiple comparisons)	No

Note. Notice that although some templates ask you to identify a research question or prediction as a simple way to help readers understand the focus of your study, you can create a pre-analysis plan even when you have no prediction about how your results will turn out (see Ledgerwood, 2018). Notice too that pre-analysis plans must be specific to be useful (for example, if the dependent measure does not specify how many items will be averaged, it is not clear whether the decision about which items to include was made before or after seeing the data).

Distinguish between Different Varieties of Preregistration and their Respective Functions

As described above, pre-analysis plans can be very useful for clearly distinguishing between data-independent and data-dependent analyses. *Preregistering* a pre-analysis plan simply means recording it in a public repository (e.g., OSF, AsPredicted.org, or socialscienceregistry.org). However, it's important to recognize that the term *preregistration* is used in different ways by different researchers both within and beyond psychology, and that these different definitions often map onto different goals or functions (Ledgerwood & Sakaluk, 2018; see also Navarro, 2019). Table 4 outlines the most common varieties of preregistration and their intended functions.

Table 4.

Different Definitions of Preregistration and Their Intended Purpose

Definition	**Goal**
Pre-analysis plan	Distinguish data-independent vs. data-dependent analyses; constraining unintended Type I error inflation
Write down as much information as you can about your study before you conduct it	Transparency: Someone else can check what you said you planned ahead of time against what you actually wrote down
Record your theoretical predictions	Theory falsification
Record your intuitive predictions	Figure out how good you are personally at guessing a study's outcome
Record the existence of your study in a centralized, searchable repository	Combat publication bias
Registered report	All of the above plus reviewer objectivity (reviewers can evaluate the methods and planned analyses without being biased by the whether the results fit their intuitions or theories)

Researchers in psychology often use the term "preregistration" to mean a pre-analysis plan, and advocate using this type of preregistration to reduce unintended Type I error inflation and help researchers correctly calibrate their level of confidence or uncertainty about a given set of results (e.g., Nosek, Ebersole, DeHaven, & Mellor, 2018a; Simmons, Nelson, & Simonsohn, 2017). But researchers in psychology and other disciplines also use the term "preregistration" to mean other practices that do *not* influence Type I error (although they serve other important functions). Distinguishing between different varieties or elements of preregistration and thinking carefully about their intended purpose (and whether a given preregistration successfully achieves that purpose) is crucial if we want to correctly calibrate our confidence in a given set of results.

For example, researchers across disciplines sometimes use the term "preregistration" to mean a peer-reviewed registered report, where a study's methods and planned analyses are peer reviewed before the study is conducted; in such cases, the decision about whether to publish the study is made independently from the study's results (e.g., Chambers, Munafo, et al., 2013). This type of preregistration can help constrain Type I error inflation (insofar as the analyses are specified ahead of time and account for multiple comparisons), while also achieving other goals like combatting publication bias (because the decision about whether to publish the study does not depend on the direction of the results). However, the reverse is not true: A pre-analysis plan by itself does not typically combat publication bias (primarily because in psychology, such plans are not posted in a public, centralized, easily searchable repository).

Similarly, researchers often talk about preregistration as involving recording a directional prediction before conducting a study (e.g., Nosek et al., 2018a), which can be useful for theory falsification. However, writing down one's predictions ahead of time does not influence Type I error: The probability of a given result occurring by chance does not change depending on whether a researcher correctly predicted it ahead of time (Ledgerwood, 2018). Thus, a researcher who records their predictions ahead of time without also specifying a careful pre-analysis plan runs the risk of unintended Type I error inflation (Nosek et al., 2018b).

Thus, in order to correctly calibrate our confidence in a given study's results, we need to know more than whether or not a study was "preregistered"—we need to ask *how* a study was preregistered. Did the preregistration contain a careful and complete pre-analysis plan that fully constrained flexibility in dataset construction and analysis decisions? Did the plan successfully account for multiple comparisons? And did the researcher exactly follow the plan for all analyses described as data-independent? Thinking carefully and critically about preregistration will help you identify which of the goals (if any) listed in Table 4 have been achieved by a given study, and whether you should be more or less confident in that study’s conclusions.

Table 5.

Summary of recommendations

When planning a study:

1) Assess and maximize statistical power
   - Conduct an *a priori* power analysis.
   - Identify the biggest sample size worth collecting.
   - Use sequential analysis to conserve resources when possible.
   - Consider multiple approaches to boosting power.
2) Avoid unintended or invisible Type I error inflation
   - Clearly distinguish between data-dependent and data-independent analyses with a pre-analysis plan.
   - Distinguish between different varieties of preregistration and their respective functions. If your goal is to avoid Type I error inflation, preregister a pre-analysis plan that clearly constrains researcher degrees of freedom for any planned, data-independent analyses.

When evaluating a study that has already been conducted:

1) Assess the likely power of a given statistical test
   - Conduct a sensitivity analysis to assess power to detect a range of effect sizes.
   - *Don’t* calculate “post-hoc” or “observed” power.
2) Assess the likely Type I error rate
   - Understand how data-dependent decisions inflate the Type I Error rate.
   - Look for a pre-analysis plan. Is it clear which analyses were data-independent? Are researcher decisions well constrained or flexible? Does the pre-analysis plan account for multiple comparisons and are the analyses appropriate?

Conclusion

In this tutorial, we have discussed a number of strategies that you can use to calibrate the confidence you have in the results of your own studies as well as studies from other researchers.

These strategies address typical issues researchers face when they try to assess the likely statistical power and Type I error rate of a given study. By improving our ability to gauge statistical power and Type I error, we can distinguish between study results that provide relatively strong building blocks for our research programs (those with high positive predictive value, as illustrated on the right side of Figure 1) and study results that provide more tentative evidence that needs to be replicated before we build on it (those with low positive predictive value, as illustrated on the left side of Figure 1). To help us get a better sense of the power of a study, we can develop good effect size intuitions; conduct *a priori* power analyses when we are in the planning phase of a project; and conduct sensitivity analyses when data has already been collected. To help us get a better sense of the Type I error rate of a study, we can clearly distinguish between data-dependent (exploratory) and data-independent (confirmatory) analyses using a pre-analysis plan; think critically about different varieties of preregistration; and evaluate whether a given preregistration successfully achieves its desired function(s). Together, these strategies can help improve our research methods as scientists, allowing us to maximize what we learn from the work that we do.

References

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*, 1547-1562.

Bakker, M., van Dijk, A., Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554.

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

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.

Chaiken, S., & Ledgerwood, A. (2011). A theory of heuristic and systematic information processing. *Handbook of theories of social psychology: Volume one* (pp. 246-166).

Chambers, C., & Munafo, M. (2013, June 5). Trust in science would be improved by study pre-registration. *The Guardian*. Retrieved from https://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration

Chartier, C., Kline, M., McCarthy, R., Nuijten, M., Dunleavy, D. J., & Ledgerwood, A. (2018, November 30). The cooperative revolution is making psychological science better. *APS Observer.* Retrieved from https://www.psychologicalscience.org/observer/the-cooperative-revolution-is-making-psychological-science-better

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., & Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review, 23*, 640-647.

Erdfelder, E., Faul, F., & Buchner, A. (2005). Power analysis for categorical methods. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1565-1570). CHicester, U.K.: Wiley.

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.

Giner-Sorolla, R. (2018, January 24). Powering Your Interaction [blog post]. Retrieved from https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641-651.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, *83*, 314-320.

Hoenig, J. M., Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*, 19-24.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532.

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*, 9.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*, 701-710.

Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*, 278-292.

Le, B., Dove, N. L., Agnew, C. R., Korn, M. S., & Mutso, A. A. (2010). Predicting nonmarital romantic relationship dissolution: A meta-analytic synthesis. *Personal Relationships, 17*(3), 377-390.

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*, *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., & Sakaluk, J. (2018, February). Preregistration, actually: How can (and should) researchers use preregistration pluralistically? Talk presented at the Society for Improving Psychological Science preconference, held before the annual conference of the Society for Personality and Social Psychology, Portland, OR.

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., Soderberg, C. K., & Sparks, J. (2017). Designing a study to maximize informational value. In M. C. Makel & J. A. Plucker (Eds.), *Toward a More Perfect Psychology: Improving Trust, Accuracy, and Transparency in Research* (pp. 33-58). Washington, DC: American Psychological Association.

McNutt, M. (2014). Journals unite for reproducibility. *Science, 346*, 679.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11*, 730-749.

Navarro, D. (2019, January 17). Prediction, pre-specification and transparency [blog post]. Retrieved from https://featuredcontent.psychonomic.org/prediction-pre-specification-and-transparency/

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018a). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115,* 2600-2606.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018b). Reply to Ledgerwood: Predictions without analysis plans are inert. *Proceedings of the National Academy of Sciences*, *115*, E10518.

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.

Open Science Collaboration. (2015). *Science*, 28.

Pashler, H., & Wagenmakers, E-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*, 528-530.

Proschan, M. A., Lan, K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. Springer.

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.

Rivers, A. M., & Sherman, J. (2018, January 19). Experimental Design and the Reliability of Priming Effects: Reconsidering the "Train Wreck". *PsyArXiv*.

Rosenthal R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*, 293-304.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609-612.

Simons, D. J. (2018). Introducing advances in methods and practices in psychological science. *Advances in Methods and Practices in Psychological Science, 1*, 3-6.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366.

Simmons, J., & Nelson, L., & Simonsohn , U. (2017, November 6). How to Properly Preregister A Study [blog post]. Retrieved from http://datacolada.org/64

Sparks, J., & Ledgerwood, A. (2017). When good is stickier than bad: Understanding gain/loss asymmetries in sequential framing effects. *Journal of Experimental Psychology: General, 146*, 1086-1105.

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*, 886-899.

Spellman, B., Gilbert, E. A., & Corker, K. S. (2017). Open Science: What, Why, and How. *PsyArXiv*.

Wang, Y. A., & Eastwick, P. W. (in press). Solutions to the problems of incremental validity testing in relationship science. *Personal Relationships*.

Wang, Y. A., & Rhemtulla, M. (in press). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial.

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.

Westfall, J. (2016). PANGEA [computer program]. Retrieved from https://jakewestfall.shinyapps.io/pangea/

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS one, 11,* e0152719.

Endnotes

1. The positive predictive value of a study also depends on the base rate of true effects being tested in the population. For the examples we provide in this section, we assume this base rate of true effects being tested is 50%. When it is higher (e.g., when a researcher tests incremental questions in a very well-established research literature where a hypothesis is quite likely to be true), the positive predictive value of a study will be higher. When it is lower (e.g., when a researcher tests a bold new idea in a new research literature, or postulates a counterintuitive effect), the positive predictive value of a study will be lower. The formula for computing positive predictive value is PPV = (Power*R) / (Power*R + α), where R is the odds of a found effect indeed being non-null among the effects being tested, depending on the base rate of true effects (see Button et. al., 2013; Pashler & Harris, 2012; Ioannidis, 2005 for fuller discussions of positive predictive value).
2. Note that although a Type I error rate of .30 is quite far from most researchers' desired Type I error rate of .05, it is not unrealistic to suggest that the actual Type I error rate of a study can be considerably higher than the nominal alpha rate (typically *p* < .05). We return to this issue in Part II of this tutorial (see also Simmons, Nelson, & Simonsohn, 2011, for a vivid demonstration, and Gelman & Loken, 2014, for an in-depth discussion).

Chapter 2

Disentangling Bias on Implicit and Explicit Measures from Responses to Exemplars and Categories

Abstract

Implicit measures often show dissociations from explicit measures, including low correlations, distinct antecedents, and distinct behavioral correlates. Interpretations of these dissociations referring to measurement types presuppose that the distinction between implicit and explicit measures is not confounded with other content-related differences. However, in research on social biases, explicit measures often focus on responses to abstract social categories, whereas implicit measures include presentations of specific exemplars of those categories. The current work addressed this confound by investigating associations between implicit and explicit measures of exemplar and category evaluations, respectively. Experiments 1 and 2 tested whether implicit and explicit measures of racial bias show stronger associations when they correspond in terms of their focus on exemplars or categories. Experiments 3 and 4 investigated whether presumed moderators of implicit-explicit relations qualify the association between measures that focus on categories versus exemplars, rather than implicit and explicit measures *per se.* Results suggest that, while the exemplar-category distinction plays a negligible role in the Affect Misattribution Procedure, previously obtained dissociations obtained with the Implicit Association Test may be at least partly driven by the exemplar-category confound.

Keywords: categories; exemplars; implicit measures; prejudice; racial bias

Arguably, one of the most impactful inventions within the field of psychology during the last three decades has been the development of implicit measures (for reviews, see Gawronski et al., 2020; Greenwald & Lai, 2020). A central feature of implicit measures is that they allow researchers to infer evaluative responses to attitudinal stimuli from objective performance indicators (e.g., speed, accuracy) rather than direct self-reports. Implicit measures have gained popularity partly because they often show dissociations from explicit self-report measures, including low correlations between the two kinds of measures (Nosek, 2005), distinct antecedents (Gawronski & Bodenhausen, 2006), and distinct behavioral correlates (Friese et al., 2008). In research on prejudice and stereotyping, these dissociations are often interpreted as evidence for the idea that implicit measures capture social biases that people are unwilling or unable to report (Greenwald & Banaji, 1995; Fazio et al., 1995).

Although claims that implicit measures provide a window into unconscious biases are controversial (see Corneille & Hütter, 2020; Fazio & Olson, 2003; Gawronski et al., 2022a; Hahn et al., 2014), many researchers agree that responses on implicit measures tend to be unintentional and difficult to control (e.g., De Houwer & Boddez, 2022; Gawronski et al., 2022b; Melnikoff & Kurdi, 2022; Olson & Gill, 2022; Ratliff & Smith, 2022). Thus, researchers have assumed that dissociations between responses on implicit measures (presumably unintentional and difficult to control) and responses on explicit measures (presumably intentional and easy to control) can be interpreted in terms of the two aspects of automaticity: intentionality and controllability (Bargh, 1994; Moors & De Houwer, 2006). Yet, such an interpretation presupposes that the distinction between implicit and explicit measures is not confounded with other important differences between measures.

In the current work, we identify an important and so far unstudied confound between many implicit and explicit measures of racial bias: the confound between type of measure and type of attitudinal stimuli. Whereas explicit measures of racial bias typically focus on responses to *abstract social categories*, most implicit measures include presentations of *specific exemplars* of those categories (e.g., Fazio et al., 1995; Greenwald et al., 1998; Payne et al., 2005; for a notable exception, see Wittenbrink et al., 1997). Pulling apart and taking seriously the distinction between explicit and implicit measures on the one hand, and the distinction between categories and exemplars on the other, is essential for gaining new insights into the causes and consequences of social biases, with important implications for both theory and practice.

Measurement and Construct

Three implicit measures stand out in terms of the frequency with which they have been used in research on racial bias: the Evaluative Priming Task (EPT; Fazio et al., 1995), the Implicit Association Test (IAT; Greenwald et al., 1998), and the Affect Misattribution Procedure (AMP; Payne et al., 2005). In a typical EPT, participants are briefly presented with an attitudinal prime stimulus (e.g., a White or Black face), which is followed by a positive or negative target word. Participants' task is to indicate as quickly as possible whether the target word is positive or negative. The idea underlying the EPT is that quick and accurate responses to the target words should be facilitated when they are evaluatively congruent with participants' attitude toward the prime stimulus. In contrast, quick and accurate responses to the target words should be impaired when they are evaluatively incongruent with participants' attitude toward the prime stimulus (Fazio, 2001). For example, if a person holds more favorable attitudes toward White people than Black people, this person should be faster and more accurate in identifying the valence of positive words when the person has been primed with an image of a White person compared to when they have been primed with an image of a Black person. Conversely, the person should be slower and less accurate in identifying the valence of negative words when they have been primed with an image of a White person compared to when they have been primed with an image of a Black person.

The most prominent implicit measure in research on racial bias is the IAT (Greenwald et al., 1998). In the critical blocks of the IAT, participants are asked to complete two binary categorization tasks that are combined in a manner that is either congruent or incongruent with the to-be-measured attitude. For example, in the commonly used race IAT, participants may be asked to categorize pictures of Black and White faces in terms of their race and positive and negative words in terms of their valence. In one critical block of the task, participants are asked to press one response key for Black faces and negative words, and another response key for White faces and positive words (i.e., prejudice-congruent block). In the other critical block, participants are asked to complete the same categorization tasks with a reversed key assignment for the faces, such that they have to press one response key for White faces and negative words, and the other response key for Black faces and positive words (i.e., prejudice-incongruent block). The basic idea underlying the IAT is that responses in the task should be facilitated when two mentally associated concepts are mapped onto the same response key. For example, a person who has more favorable attitudes toward White people than Black people should show faster and more accurate responses when White faces share the same response key with positive words and when Black faces share the same response key with negative words, compared with the reversed mapping.

The AMP was designed to combine the structural advantages of the EPT with the superior psychometric properties of the IAT (Payne et al., 2005). Two central differences of the

AMP are that (1) the target stimuli in the AMP are ambiguous and (2) participants are asked to report their subjective evaluations of the targets. The basic idea is that participants may misattribute the affective feelings elicited by the prime stimuli to the neutral targets, and therefore judge the targets more favorably when they were primed with a positive stimulus than when they were primed with a negative stimulus (for a review, see Payne & Lundberg, 2014). For example, in an AMP to measure racial attitudes, participants may be asked to indicate whether they find Chinese ideographs visually more pleasant or visually less pleasant than average after being primed with pictures of Black versus White faces. A preference for White over Black people would be indicated by a tendency to evaluate the Chinese ideographs more favorably when the ideographs followed the presentation of a White face than when they followed the presentation of a Black face.

The development of implicit measures paved the way for considerable theory and research, which can be characterized by three core themes (for a review, see Gawronski et al., 2020). First, research on the relation between implicit and explicit measures revealed that correlations between the two are often quite low (for meta-analyses, see Cameron et al., 2012; Hofmann, Gawronski, et al., 2005). Second, research on the antecedents of responses on implicit and explicit measures suggests that they may differ in their sensitivity to different kinds of information (for a review, see Gawronski & Bodenhausen, 2006). Finally, research on behavioral correlates of implicit and explicit measures suggests that they may predict different kinds of behaviors (e.g., unintentional vs. intentional behavior) and behavior under different contextual conditions (e.g., high vs. low cognitive load; for a review, see Friese et al., 2008).

The available evidence for low correlations between implicit and explicit measures, distinct antecedents, and distinct behavioral correlates have led many researchers to conclude that implicit and explicit measures capture related but distinct constructs (e.g., Bar-Anan & Vianello, 2018; Nosek & Smyth, 2007). Such a conclusion would seem justified to the extent that the relevant evidence was obtained with implicit and explicit measures that have conceptually equivalent content (e.g., use the same stimuli) aside from the implicit and explicit nature of their measurement approaches. However, a closer inspection of the literature suggests that this is not always the case, with research on racial bias standing out as a particularly problematic area (see Axt, 2018; Gawronski, 2019; Payne et al., 2008). To the extent that an implicit measure has little or no content correspondence with an explicit measure, their relation can be expected to be low for important but often overlooked methodological reasons (Ajzen & Fishbein, 1977). In such cases, it would be premature to interpret their weak relation as evidence for the hypothesis that implicit and explicit measures capture distinct constructs (e.g., Bar-Anan & Vianello, 2018; Nosek & Smyth, 2007). Similarly, if type of measure is confounded with different contents, any finding suggesting distinct antecedents or distinct behavioral correlates remains ambiguous, because the obtained dissociation could be due to either (1) the implicit versus explicit nature of the measures or (2) the different contents of the two measures.

Distinguishing between Exemplars and Categories

In our view, one of the most important—and as yet unstudied—confounds between implicit and explicit measures of bias in previous research is that they typically have differed in terms of the relevant attitudinal stimuli presented to participants. Specifically, implicit versus explicit measures of bias often differ in the extent to which they involve responses to specific *exemplars* or abstract *categories* (see Gawronski, 2019). A common practice in bias research using implicit and explicit measures is to use specific images of exemplars as attitudinal stimuli in the implicit measure (e.g., images of Black and White faces in the EPT, the IAT, and the

AMP) and to use abstract category labels in the explicit measure (e.g., the categories *Black people* and *White people* in feeling thermometer ratings or semantic differentials). Although it seems reasonable to assume that a person's responses to exemplars of a given category are related to that person's responses to the category in general, evaluations of exemplars and evaluations of categories are conceptually distinct. Whereas evaluations of exemplars may reflect people's experienced responses to concrete targets, evaluations of categories may reflect people's abstract ideas about their likes and dislikes (Eastwick et al., 2019; Ledgerwood et al., 2018).

More broadly, although bias research using implicit measures has tended to assume that categories and exemplars are interchangeable (for notable exceptions, see Cooley & Payne, 2019; Livingston & Brewer, 2002; Olson & Fazio, 2003), multiple cognitive and social psychological literatures have long treated them as distinct constructs. Historically, categories and exemplars have animated very different kinds of models of memory and social judgment (Alba & Hasher, 1983; Brooks, 1987; Smith & Zárate, 1992). With respect to the former, schema-based models focus on general, abstract knowledge (e.g., knowledge about categories) that people apply in a top-down fashion (Smith, 1998). With respect to the latter, exemplar-based models posit that specific past experiences with particular exemplars can influence memory and judgment via relatively detailed, concrete cognitive representations (Smith & Zárate, 1992).

More recently, construal level theory has highlighted the distinction between categories and exemplars in terms of level of abstraction. Whereas categories capture the superordinate, prototypical, and general features of a set of objects, exemplars instantiate specific, concrete instances that vary in terms of their subordinate and peripheral details (Ledgerwood et al., 2010; McCrea et al., 2012; Soderberg et al., 2015; Trope & Liberman, 2010). Consistent with this notion, research suggests that thinking about categories versus exemplars can have different consequences for a range of cognitive processes and social judgments related to abstraction (Fujita & Han, 2009; Hansen & Trope, 2013; Ledgerwood et al., 2010; Wakslak & Trope, 2009). In fact, evaluations reflecting more concrete versus abstract responses often show low correlations, have different antecedents, and predict different downstream consequences—even if all the measures derive from explicit self-reports (da Silva Frost et al., 2022; Ledgerwood & Wang, 2018).[1]

The Current Research

In sum, previous research has concluded that implicit and explicit measures capture related but distinct constructs based on evidence that implicit and explicit measures have low correlations, distinct antecedents, and distinct behavioral correlates. However, our reasoning above suggests that such conclusions are premature in the absence of more compelling evidence that the observed dissociations indeed reflect the presumed differences between implicit and explicit measures rather than differences in responses to exemplars versus categories. To the extent that research disentangling the confound between type of measure and content of measure suggests that distinct constructs underlie responses on implicit and explicit measures, such evidence would provide a stronger basis for conclusions about unique properties of bias on implicit measures and its potential significance for social discrimination in real-world contexts. Conversely, if such research confirmed a distinct role of exemplar versus category evaluations

[1] Like Ajzen and Fishbein (1977, 2005), we argue that evaluative responses may seem to have weak correspondence in cases where researchers specify the object of evaluation in mismatching ways. For example, according to Ajzen and Fishbein, a general attitude (e.g., attitudes toward environmentalism) might fail to predict a specific behavior (e.g., voting on a city ordinance that would require composting), because the two attitude objects are not specified at the same level (see also Ledgerwood & Trope, 2010). It is worth noting, however, that the distinction we make here between evaluations of categories (e.g., a person's evaluation of the category *Black people*) and evaluations of exemplars (e.g., a person's average evaluation of a set of Black faces) is different in that, in Ajzen and Fishbein's (1977) work, these two measures of evaluations were actually treated as interchangeable methods for assessing the same general attitude construct (see Ledgerwood et al., 2018).

that is independent of the distinction between implicit and explicit measures, the obtained evidence would necessitate a significant correction of previous interpretations, offer a novel conceptual foundation for future research, and suggest different interventions to reduce social discrimination in real-world contexts.

In the current work, we addressed these issues for racial-bias applications of the AMP and the IAT. Our focus on the two measures was based on three considerations. First, they are the two implicit measures with the highest internal consistencies (Gawronski & De Houwer, 2014; Greenwald & Lai, 2020), which is methodologically imperative for the correlational designs of the current studies (see Koppehele-Gossel et al., 2020). Second, the two measures are based on different underlying mechanisms (Gawronski & De Houwer, 2014; Nosek et al., 2011), which is essential for testing the generality or specificity of our results (see Gawronski et al., 2008). Third, the two measures have unique strengths and weaknesses, allowing us to complement the weaknesses of one measure with the strengths of the other (and vice versa).

We sought to design all studies so that they would be informative regardless of the way the results turned out. The main goal of Experiments 1 and 2 was to test whether implicit and explicit measures of bias show stronger associations when they correspond (versus do not correspond) in terms of their focus on exemplars or categories. Toward this end, Experiments 1 and 2 assessed the correlations between implicit and explicit measures of racial bias that either mismatch (i.e., one exemplar-focused and one category-focused) or match (i.e., both exemplar-focused or both category-focused) in terms of their contents. In Experiment 1, participants completed an AMP measure of either (1) responses to specific Black and White faces (i.e., implicit exemplar measure) or (2) responses to category labels like *Black people* and *White people* (i.e., implicit category measure). In addition, participants completed a feeling thermometer measure of either (1) responses to specific Black and White faces (i.e., explicit exemplar measure) or (2) responses to category labels (i.e., explicit category measure). We tested whether implicit-explicit relations are stronger when they correspond in terms of their focus on exemplars or categories, compared to when they do not.

In Experiment 2, participants completed a typical race IAT, a feeling thermometer with exemplars, and a feeling thermometer with categories. The IAT is different from many other implicit measures, in that it includes specific faces as target stimuli and category labels for the response options. Thus, variance in the IAT might be jointly driven by responses to exemplars and categories. Based on these considerations, we tested whether the association between IAT scores and each of the two explicit measures remains significant when controlling for the respective other explicit measure. A significant association between IAT scores and a category-level explicit measure, controlling for an exemplar-level explicit measure, would suggest that responses to categories play a role in IAT responses. Conversely, a significant association between IAT scores and an exemplar-level explicit measure, controlling for a category-level explicit measure, would suggest that responses to exemplars play a role in IAT responses.

Expanding on the findings of first two studies, Experiments 3 and 4 investigated whether presumed moderators of implicit-explicit relations qualify the correspondence between measures that focus on categories versus exemplars, rather than implicit and explicit measures *per se*. Past research has found that individuals who report a strong motivation to control prejudiced reactions show weaker relations between implicit and explicit measures of bias, compared to individuals who report a weak motivation to control prejudiced reactions (e.g., Degner & Wentura, 2008; Dunton & Fazio, 1997; Fazio et al., 1995; Gawronski et al., 2003; Payne et al., 2005). We suspect that this pattern is at least partly driven by an attenuating effect of motivation to control prejudice on the correspondence between category-level and exemplar-level evaluations. This depressed correspondence could occur for two reasons: (1) because the motivation to control prejudice operates at the level of categories rather than specific exemplars (i.e., people who want to control prejudice modulate their abstract evaluations of categories but not their experienced responses to specific exemplars), or (2) because the motivation to control prejudice interferes with the translation of experienced responses to specific exemplars into abstract category evaluations (i.e., people who want to control prejudice avoid generalizing inferences from experienced responses to specific exemplars to abstract evaluations of overall categories).

To test these hypotheses, participants in Experiments 3 and 4 completed Dunton and Fazio's (1997) Motivation to Control Prejudice Reactions Scale (MCPRS) and a typical feeling thermometer with category labels. In Experiment 3, half of the participants additionally completed a typical exemplar-focused AMP with Black and White faces, while the remaining half completed an exemplar-focused feeling thermometer with the same Black and White faces. We tested the prediction that, regardless of whether the exemplar-focused measure is implicit or explicit, the relation between the exemplar-focused measure and the category-focused measure decreases as MCPRS scores increase. Participants in Experiment 4 completed a typical IAT (with faces as target stimuli and category labels for the response options), an exemplar-focused feeling thermometer, and a typical feeling thermometer with category labels. We tested the prediction that MCPRS scores moderate not only the relation between racial bias on the IAT and the explicit category-focused measure (replicating earlier findings), but also the relation between the explicit exemplar-focused measure and the explicit category-focused measure.

Open Practices

For all studies, we report how we determined our sample sizes, all data exclusions, all manipulations, and all measures. The data, analysis codes, and research materials for all studies are available at https://osf.io/d9q3g/?view_only=083207bb96744d86a95afc8f80d7eff2. For all studies, we preregistered our analysis plan, including target sample size and exclusion criteria. Hyperlinks to the preregistrations are reported in the Methods sections of each study.

Experiment 1

The main goal of Experiment 1 was to test if the correlations between explicit and implicit measures of racial bias are higher when they correspond (versus do not correspond) in terms of their contents (exemplar vs. category). To this end, each participant completed one AMP and one feeling-thermometer measure, either of which could be an exemplar measure or a category measure. If correspondence in terms of content is indeed relevant for interpreting relations between the AMP and explicit measures, correlations between measures should be higher when participants completed two measures with matching content (i.e., both exemplar or both category; see A and C in Figure 3) than when they completed two measures with mismatching content (i.e., one exemplar and one category; see B and D in Figure 3).

Method

Preregistration

The analysis plan was publicly preregistered at https://aspredicted.org/ZUC_ABW.

Participants

We preregistered a target sample size of 1250, so that after an estimated 50 exclusions, we would have 300 participants per condition. We based this target sample on an a priori power analysis in G*Power (Faul et al., 2007), which indicated that we would need 300 participants per condition in order to achieve 80% power to detect a difference between correlations of *q* = .23 between any two of the four conditions. Participants were MTurk workers over the age of 18 who completed the experiment on Inquisit. The raw dataset has 1252 rows. Out of these, 1 is missing data on key measures, 2 are incompletes and 2 are duplicates. We preregistered that we would exclude participants who either (1) reported knowing the meaning of the ideographs in the AMP (answering *yes* to a *yes*/*no* question), or (2) provided a nonsensical response to a Winograd-like attention check, or (3) pressed the same key on all trials of the AMP. After excluding the 126 participants who met one or more of these exclusion criteria (*n* = 87 reported knowing the ideographs, *n* = 33 failed the Winograd-like attention check, and *n* = 7 pressed the same key), analyses were conducted on the remaining 1121participants (624 women, 486 men, and 11 people who chose a different option; M_{age} = 39.3, *SD* = 12.8; 80.8% White, 8% Black, 8.1% Hispanic or Latino/a/e, 3.5% Black and White Biracial, 2.1% South Asian, 3.7% East Asian, and 1.7% a different identity). For the smallest cell sizes in the four experimental conditions, the final sample provides 80% power to detect a difference between correlations of *q* = .25.

Procedure and Materials

Participants were randomly assigned to one of four between-subjects conditions. In each condition, participants completed one explicit and one implicit measure (the order of the two measures was counterbalanced). The four conditions were: (1) explicit exemplar and implicit exemplar; (2) explicit exemplar and implicit category; (3) explicit category and implicit category; and (4) explicit category and implicit exemplar (see Figure 3). For the category measures, we used the labels *African American*, *European American*, *Black people*, and *White people*. For the exemplar measures, we used 20 faces of Black and White women and men (5

Black women, 5 White women, 5 Black men, and 5 White men). Afterwards, participants completed an attention check and demographic questions.[2]

Explicit exemplar measure. The explicit exemplar measure was a feeling thermometer in which participants rated their feelings toward each face on a 7-point scale ranging from 1 (*very negative*) to 7 (*very positive*). The faces were presented one at a time in random order. To create an index of explicit preference for White over Black exemplars, we computed the average explicit ratings of the ten White faces and the average explicit ratings of the ten Black faces, and then subtracted the average ratings of Black faces from the average ratings of White faces.

Explicit category measure. The explicit category measure consisted of six feeling thermometer ratings. On four of the six items, participants rated their feelings toward each of the four categories (i.e., African American, European American, Black people, and White people) on 5-point scales ranging from 1 (*very negative*) to 5 (*very positive*). On the remaining two items, participants rated their relative preference for one category over the other (i.e., African American vs. European American; Black people vs. White people) on 9-point scales ranging -4 (*strongly prefer African Americans / Black people over European Americans / White people*) to +4 (*strongly prefer European Americans / White people over African Americans / Black people*). To create an index of explicit preference for White over Black categories, we first computed difference scores for each of the two pairs of single-category feeling thermometers, and then averaged the resulting two difference scores with the two relative preference ratings (all coded in the same direction).

Implicit exemplar measure. The implicit exemplar measure was an AMP measuring responses to the same faces used in the explicit exemplar measure described above. On each trial

[2] In all studies, we also included a funnel debriefing at the end to probe for suspicion, but this measure was not part of the preregistered exclusion criteria.

of the task, participants were first presented with a fixation cross for 500 ms, which was replaced by a picture of either a Black or a White face for 75 ms. The presentation of the face prime was followed by a blank screen for 125 ms, after which a Chinese ideograph appeared for 100 ms. The Chinese ideograph was then replaced by a black-and-white pattern mask, and participants had to indicate whether they considered the Chinese ideograph as more pleasant or less pleasant than the average Chinese ideograph. The pattern mask remained on the screen until participants gave their response. Participants were asked to press a right-hand key (*I*) if they considered the Chinese ideograph as more pleasant than average, and a left-hand key (*E*) if they considered the Chinese ideograph as less pleasant than average. Following the instructions employed by Payne et al. (2005), participants were told that the pictures can sometimes bias people's responses to the Chinese ideographs, and that they should try their absolute best not to let the pictures influence their judgments of the Chinese ideographs. Each of the 20 face primes was presented twice, summing up to a total of 40 trials. As target stimuli, we used 40 distinct Chinese ideographs from Payne et al. (2005). Order of trials was randomized for each participant. To create an index of implicit preference for White over Black exemplars, we calculated the proportion of *pleasant* responses to target stimuli for each prime type (i.e., Black face vs. White face), and then subtracted the proportion score for Black face primes from the proportion score for White face primes.

Implicit category measure. The implicit category measure was an AMP measuring responses to the same four category labels used in the explicit category measure described above. The procedural details were identical to the AMP measuring responses to exemplars, the only difference being that we used the four category labels as primes (i.e., *African American*, *European American*, *Black people*, and *White people*) instead of Black and White faces. Each of the category labels was presented ten times across 40 trials. To create an index of implicit preference for White over Black categories, we calculated the proportion of *pleasant* responses to target stimuli for each prime type, and then subtracted the proportion score for Black category primes from the proportion score for White category primes.

Knowledge about the meaning of the ideographs. In line with previous AMP practices (e.g., Gawronski & Ye, 2014), we asked participants *Do you know the meaning of the Chinese ideographs we showed you in the concentration task?* (Yes/No) in the demographics section at the end of the study.

Winograd-like attention check. To identify bots and inattentive participants, we used a Winograd-like attention check, which is part of the standard operating procedures in the first and second authors' lab. This check involves text interpretation based on the structure of a Winograd schema (used to assess human-like reasoning; Levesque et al., 2012). To this end, participants read the following story: *An elderly man had the dream of watching the American female soccer team playing for their country. His grandson bought him a ticket to travel to Brazil during the 2016 Olympics as a gift. When he woke up on the day of his birthday and received the ticket, he cried of happiness. Thinking about the game, he hoped to be able to see legends such as Carli Lloyd, Megan Rapinoe, and Marta up close.* Next, they answered two open-ended questions about the story (*Who got a birthday gift?* and *What does the man expect for Summer 2016?*). Participants were excluded if they gave nonsensical answers (e.g., "unhappily") to either question, as coded by a researcher without knowledge of the results.

Results

Descriptive statistics of the four measures are presented in Table 6. If measurement content affects correspondence, correlations between measures with matching content (A and C in Figure 3) should be higher than the correlations between measures with mismatching content (B and D in Figure 3). Following our preregistered analysis plan, we conducted *z*-score tests to examine if there were significant differences between the four pairs of correlations (see Table 7). As predicted, the comparison between B and C showed a significant difference such that the matching content correlation C was higher than the mismatching content correlation B ($z = 2.37$, $p = .018$). Although the pattern of differences were in the expected direction for the other three comparisons (i.e., matching content > mismatching content), we did not find a significant difference when comparing A and B ($z = 1.22$, $p = .223$), C and D ($z = 1.89$, $p = .058$), and A and D ($z = 0.71$, $p = .478$).

Discussion

In this experiment, we tested if the correspondence between explicit and implicit measures was higher when the measures corresponded (versus did not correspond) in terms of their focus on exemplars or categories. Although the pattern of differences was in the expected direction, the differences were fairly small and only one comparison reached statistical significance. Overall, these results suggest that the distinction between category and exemplar measures—although important to consider for methodological reasons—does not seem particularly consequential for correlations between the AMP and explicit measures. By disentangling the confound between type of measure and content of measure, the current experiment permits greater confidence that distinct constructs underlie responses on the AMP and feeling-thermometer measures, even if they differ in terms of their focus on specific exemplars versus abstract categories.

Experiment 2

In Experiment 2, we investigated whether the correspondence between racial bias on the IAT and feeling-thermometer measures depends on the content of the explicit measure (categories vs. exemplars). Different from the AMP, the standard race IAT includes both exemplars (i.e., faces as target stimuli) and categories (i.e., category labels for the response options). Thus, variance in IAT scores might reflect a mix of responses to exemplars and responses to categories. Based on these considerations, we tested whether (1) the association between IAT scores and the explicit category measure remains significant when controlling for the explicit exemplar measure, and (2) the association between IAT scores and the explicit exemplar measure remains significant when controlling for the explicit category measure. A significant association between IAT scores and a category-level explicit measure, controlling for an exemplar-level explicit measure, would suggest that responses to categories play a role in IAT responses. Conversely, a significant association between IAT scores and an exemplar-level explicit measure, controlling for a category-level explicit measure, would suggest that responses to exemplars play a role in IAT responses.

Method

Preregistration

The analysis plan was publicly preregistered at https://osf.io/d9q3g/?view_only=083207bb96744d86a95afc8f80d7eff2 .

Participants

We conducted an a priori power analysis using the Shiny App pwrSEM, which is based on Monte Carlo simulations (Wang & Rhemtulla, 2020). We aimed for at least 80% power to detect a significant association (with $\alpha = .05$) in two partial regressions of the explicit measures of racial bias on the IAT, as well as model misspecification based on the procedure described by MacCallum et al. (1996). Based on extant reviews (see Gawronski & De Houwer, 2014; Greenwald & Lai, 2020), we assumed the reliability of the explicit category-focused measure to be α = .90, of the explicit exemplar-focused measure to be α = .80 and of the IAT to be α = .70. Moreover, based on related data reported by Gawronski (2019), the power analysis was based on partial correlations of $r = .12$ between the IAT scores and each of the two explicit measures. We ran 1000 simulations (set seed = 420 and 7) and the app indicated that the minimum target sample size that would provide 80% power to detect all three effects was $N = 545$.[3] We anticipated an exclusion rate of approximately 8% and oversampled to a target sample size of 600, to ensure that we would have at least $N = 545$ for analysis. We preregistered that, if after exclusions and before running any analyses we had a sample size of less than 550, we would compute the exclusion rate (*e*) and collect $n = (550\text{-current } n)/(1\text{-}e)$ additional participants. We also preregistered that we would exclude participants who (1) showed latencies lower than 300 ms on 10% or more of the trials, or (2) provided a nonsensical response to a Winograd-like attention check designed to filter out bots and inattentive participants.

Participants were Prolific workers over the age of 18, currently living in the USA, fluent in English, and with 80% approval rate who completed the experiment online on Inquisit. A total of 602 rows appear in the raw dataset. Two of these participants were duplicates, and seven had an error in their subject IDs, which made us unable to aggregate their data across tasks. After excluding 8 additional participants who met one or more of our preregistered exclusion criteria ($n = 0$ showed latencies lower than 300 ms on 10% or more of the trials, and $n = 8$ failed the Winograd-like attention check), analyses were conducted on the remaining 585 participants (279

[3] The two partial regressions plus model misspecification were based on the procedure described by MacCallum et al. (1996).

women, 292 men, and 14 people who chose a different option; M_{age} = 30.8, *SD* = 11.6; 65.6% White, 8.5% Black, 12.6% Hispanic or Latino/a/e, 2.7% Black and White Biracial, 5.1% South Asian, 12.5% East Asian, 1.5% American Indian or Alaskan Native, and 3.6% a different identity). A sensitivity analysis conducted on pwrSEM with the same parameter values as above indicated that our final sample provides 83% power to detect a significant path from the explicit category measure and 86% power to detect a significant path from explicit exemplar measure.

Procedure and Materials

All participants completed a standard race IAT, an explicit category measure, and an explicit exemplar measure, followed by attention checks and the demographics questions. The order of the three measures of racial bias was counterbalanced across participants. The order of the stimuli was randomized within each of the three bias measures, the only constraint being that faces and adjectives were presented in alternating order in the IAT. We used the labels *Black people* and *White people* as category labels in the IAT and the explicit category measure; the adjectives *friendly*, *unfriendly*, *likable*, *dislikable*, *pleasant*, *unpleasant*, *nice*, *nasty*, *good*, and *bad* were used as attribute stimuli in the IAT and for the ratings in the explicity category measure. For the exemplar measure, we used the same 20 faces of Black and White women and men from Experiment 1.

Explicit exemplar measure. The explicit exemplar measure was identical to the one in Experiment 1.

Explicit category measure. To match the stimuli used in the IAT, the explicit category measure consisted of 10 adjective ratings for the categories Black and White (e.g., *White people are unpleasant*). Responses were measured with 7-point scales ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

IAT measure. The IAT included 80 trials on the compatible and incompatible blocks, respectively, using the faces of the explicit exemplar measure as target stimuli and the bipolar adjectives of the explicit category measure as attribute stimuli. In the compatible block (same direction as societal bias), participants had to press the *I* key for positive adjectives and White faces, and the *E* key for negative adjectives and Black faces. In the incompatible block, participants had to press the *I* key for positive adjectives and Black faces, and the *E* key for negative adjectives and White faces. The order of the compatible and the incompatible block was counterbalanced across participants. If participants gave an incorrect response, they had to provide the correct response before being able to proceed.

Results

Descriptive statistics of and zero-order correlations between the three measures of racial bias are presented in Table 8. All measures of racial bias showed significant positive associations at the level of zero-order correlations.

For our main analysis, we preregistered that we would use structural-equation modeling (SEM) with latent variables to test (1) the association between the exemplar-focused explicit measure and IAT scores while controlling for the category-focused explicit measure, and (2) the association between the category-focused explicit measure and IAT scores while controlling for the exemplar-focused explicit measure.

Following our preregistered analysis plan, we computed 5 parcels for the explicit exemplar measure, each reflecing the average of two White faces minus the average of two Black faces. Thus, each parcel is an index of explicit preference for White over Black exemplars. Parcel 1 consisted of White male face 1, White female face 1, Black male face 1 and Black female face 1; parcel 2 consisted of the faces numbered 2, and so on.

For the explicit category measure, we computed (1) the difference between ratings of the category White and ratings of the category Black for each of the five positive adjectives (i.e., friendly, likable, pleasant, nice, good), and (2) the difference between ratings of the category Black and ratings of the category White for each of the five negative adjectives (i.e., unfriendly, dislikable, unpleasant, nasty, bad). Next, we computed 5 parcels, each consisting of the average difference score for two adjectives that are direct opposites (e.g., friendly-unfriendly, likeable-dislikeable). Thus, each parcel is an index of explicit preference for the category White over the category Black.

For the IAT, we computed the D-score with built-in error penalty recommended by Greenwald et al. (2003). The two combined blocks were further broken down into 4 sub-blocks with 20 trials each.

The model showed good fit of the data, $\chi^2(74) = 154.25$, $p < .001$, Comparative Fit Index (CFI) = .988, Tucker-Lewis Index (TLI) = .985, Root Mean Square Error of Approximation (RMSEA) = 0.04. The path coefficient for the explicit exemplar measure predicting IAT scores was relatively large and statistically significant when controlling for the explicit category measure, b(SE) = 0.14(0.04), $p < .001$, $\beta = .31$. In contrast, the path of the explicit category measure predicting IAT scores was relatively small and not statistically significant when controlling for the explicit exemplar measure, b(SE) = 0.02(0.03), $p = .451$, $\beta = .06$ (see Figure 4).

Discussion

Although racial bias on the IAT, an explicit category measure, and an explicit exemplar measure were all positively associated at the level of zero-order correlations, an SEM using both explicit exemplar and explicit category evaluations as predictors of IAT scores revealed a significant path only for the explicit exemplar measure after controlling for the explicit category measure. The path from the explicit category measure was not significant after controlling for the explicit exemplar measure. Thus, different from the dominant assumption that the category labels in the IAT make it a measure of category evaluations (e.g., De Houwer, 2001; Fazio & Olson, 2003), the IAT behaved more like a measure of exemplar evaluations, and the association between IAT scores and the explicit category measure seems to be driven by the shared variance between the explicit category measure and the explicit exemplar measure. Together, these results suggest that, when considering relations between the IAT and explicit measures, the distinction between exemplars and categories does matter.

Experiment 3

Experiment 2 provided evidence that the focus on exemplars versus categories may indeed matter for the correspondence between explicit and implicit measures. Next, we set out to investigate whether presumed moderators of implicit-explicit relations qualify the association between measures that focus on categories versus exemplars, rather than implicit and explicit measures *per se*. For this purpose, we chose motivation to control prejudice, which has been found to moderate relations between implicit and explicit measures of bias (e.g., Degner & Wentura, 2008; Dunton & Fazio, 1997; Fazio et al., 1995; Gawronski et al., 2003; Payne et al., 2005). Specifically, participants who report a strong motivation to control prejudice have been found to show a weaker relation between implicit and explicit measures of bias compared to participants who report a weak motivation to control prejudice.

Given the findings of our first two experiments, it seemed important to test whether the obtained moderation pattern may be rooted in the different foci on categories versus exemplars rather than the implicit versus explicit nature of the measures. As we noted in the Introduction, motivation to control prejudice may moderate relations between responses to exemplars and categories, because (1) motivation to control prejudice operates at the level of categories rather than specific exemplars or (2) because motivation to control prejudice interferes with the abstraction from experienced evaluations of exemplars to categories (or both). To test this possibility, participants in Experiment 3 completed either an exemplar-focused AMP or an explicit exemplar-focused feeling thermometer with the same faces. In both conditions, participants additionally completed an explicit category measure and the MCPRS (Dunton & Fazio, 1997). We tested whether the relation between the exemplar-focused measures and the explicit category-focused measure decreases as MCPRS scores increase, and whether this pattern emerges irrespective of whether the exemplar-focused measure is implicit or explicit.

Method

Preregistration

The analysis plan was publicly preregistered at https://osf.io/d9q3g/?view_only=083207bb96744d86a95afc8f80d7eff2.

Participants

We preregistered a target sample size of 1080, as indicated by completed surveys on Prolific. We further preregistered that, if after exclusions and before running any analyses we had a sample size of less than 1080, we would compute the exclusion rate (*e*) and collect *n* = (1080-current *n*)/(1-*e*) additional participants. Participants were MTurk workers over the age of 18 who completed the experiment on Inquisit. A total of 1263 rows appear in the raw dataset, out of which 2 were duplicates and 91 had errors in their subject ID that prevented data aggregation.[4]

[4] We first collected data from 1080 participants. Out of these, 84 had a subject ID error and 23 were excluded based on our preregistered exclusion criteria, leaving us with a sample of 973. Following our preregistered recruitment plan, we set out to collect data from 118 additional participants. Due to a mistake, 181 were collected instead, out of

The exclusion criteria was to exclude participants who (1) provided a nonsensical response to a Winograd-like attention check designed to filter out bots or inattentive participants, and (2) pressed the same key on all trials of the AMP.[5] After excluding the 25 participants who met one or more of these criteria (*n* = 10 failed the Winograd-like attention check, and *n* = 15 pressed the same key), analyses were conducted on the remaining 1145 participants (575 women, 548 men and 22 people who chose a different option; M_{age} = 35.7, *SD* = 12.9; 75.2% White, 9.2% Black, 9.7% Hispanic or Latino/a/e, 2.5% Black and White Biracial, 2.7% South Asian, 6.4% East Asian and 2.7% a different identity). Using the zero-order correlations between the variables reported by Payne et al. (2005), a sensitivity power analysis conducted in the InteractionPoweR Shiny App (Baranger et al., 2023) indicated that a sample of 1145 (approximately 573 per condition) provides ~ 94% power to detect an interaction effect of *r* = .13, half the size of the interaction effect reported by Dunton and Fazio (1997).

Procedure and Materials

Participants were randomly assigned to one of two between-subjects conditions in which they completed either an exemplar-focused AMP or an explicit exemplar measure. In addition to completing one of the two exemplar measures, all participants completed the MCPRS and an explicit category measure. The order of the exemplar and category measures was counterbalanced, and after participants completed both measures, they completed the MCPRS. The explicit category measure, the exemplar-focused AMP, and the explicit exemplar measure were identical to Experiment 1. The MCPRS was directly adapted from Dunton and Fazio

which 7 had a subject ID error and 2 were excluded based on our preregistered exclusion criteria. Conclusions do not change with or without the extra participants included.

[5] Due to a mistake, we preregistered the exclusion criteria of the IAT (as in Experiment 2), instead of the exclusion criteria of the AMP (as in Experiment 1). We asked a co-author who did not know anything about the data to make the call, and they concluded that it did not make sense to follow the IAT exclusion criteria when using the AMP. The conclusions do not change if we include participants who reported knowing the meaning of the ideographs or not.

(1997). Responses to the 17 items of the MCPRS (e.g., *If I have a prejudiced thought or feeling, I keep it to myself*) were measured with 7-point scales ranging from -3 (*strongly disagree*) to +3 (*strongly agree*). Responses on the MCPRS were averaged to create an index with higher values indicating a stronger motivation to control prejudiced reactions. At the end of the study, participants completed an attention check and demographic questions.

Results

Descriptive statistics of and zero-order correlations between measures are presented in Table 9. Within each condition, MCPRS scores showed significant negative correlations with the measures of racial bias, and each racial-bias measure was posistively correlated with the other racial-bias measures.

Following our preregistered analysis plan, we conducted separate multiple-regression analyses for each condition. In the AMP condition, explicit category evaluations were regressed onto standardized MCPRS scores, standardized AMP scores, and the interaction term. The main effect of MCPRS scores was not significant, $\beta = -0.08$, $t(555) = -1.65$, $p = .099$, and the main effect of AMP scores was significant, $\beta = 0.34$, $t(555) = 7.22$, $p < .001$. Critically, the interaction term was significant, $\beta = -0.10$, $t(555) = -2.61$, $p = .009$. Replicating earlier findings, participants with a weak motivation to control prejudiced reactions revealed a stronger positive association between AMP scores and the explicit category measure than participants with a strong motivation to control prejudiced reactions (see Figure 5).

In the explicit-exemplar condition, explicit category evaluations were regressed onto standardized MCPRS scores, standardized explicit exemplar evaluations, and the interaction term. The main effect of MCPRS scores was not significant, $\beta = -0.07$, $t(582) = -1.81$, $p = .071$, and the main effect of explicit exemplar evaluations was significant, $\beta = 0.68$, $t(582) = 17.94$, $p <$

.001. Critically, the interaction term was not significant, $\beta = 0.017$, $t(582) = 0.58$, $p = .561$, indicating a positive association between the explicit exemplar measure and the explicit category measure regardless of motivation to control prejudiced reactions (see Figure 6).[6]

Discussion

Experiment 1 suggested that the exemplar-category difference may not matter for the AMP. Consistent with this conclusion, Experiment 3 suggests that the moderating effect of motivation to control prejudiced reactions is limited to the AMP and does not emerge for an explicit exemplar measure using the same stimuli. Together, these findings support the conclusion that prior evidence for dissociations between the AMP and explicit category measures may indeed be rooted in the type of measure (i.e., implicit vs. explicit) rather than the measured contents (i.e., exemplar vs. category).

Experiment 4

The goal of Experiment 4 was to investigate the same moderating effect of motivation to control prejudice for the IAT instead of the AMP. To this end, half of the participants completed a standard race IAT; the remaing half completed an explicit exemplar measure. As in Experiment 3, participants in both conditions also completed an explicit category measure and the MCPRS.

Method

Preregistration

The pre-analysis plan was publicly preregistered at https://osf.io/d9q3g/?view_only=083207bb96744d86a95afc8f80d7eff2.

[6] Although not preregistered, we also tested whether the moderating effect of MCPRS was qualified by the type of exemplar-focused measured. To this end, we regressed explicit category evaluations onto standardized MCPRS scores, standardized exemplar evaluations, dummy-coded type of exemplar measures (implicit = 1 and explicit = 0), and all interactions between the three predictors. Consistent with our conclusion that the moderating effect of MCPRS depended on the type of exemplar measure, the three-way interaction was statistically significant, $\beta = -.12$, $t(1137) = -2.44$, $p = .015$.

Participants

As in Experiment 3, we preregistered a target sample size of 1080, as indicated by completed surveys on Prolific. We further preregistered that, if after exclusions and before running any analyses we had a sample size of less than 1080, we would compute the exclusion rate (*e*) and collect n = (1080-current n)/(1-e) additional participants. Participants were Prolific workers over the age of 18 completed the experiment on Inquisit.[7] A total of 1141 rows appear in the raw dataset. Out of these, 30 were duplicates, 2 were incomplete and 20 had a subject ID error so that we could not aggregate their data. We preregistered that we would exclude participants who (1) responded with a latency of less than 300 ms on 10% or more of the trials or (2) provided a nonsensical response to a Winograd-like attention check designed to filter out bots or inattentive participants. After excluding the 10 participants who met one or more of these criteria (n = 4 had 10% of their trials or more with latency of less than 300 ms and n = 6 failed the Winograd-like attention check), analyses were conducted on the remaining 1079 participants (613 women, 431 men and 35 people who chose a different option; M_{age} = 39.9, SD = 14.1; 78.4% White, 6.6% Black, 9.8% Hispanic or Latino/a/e, 3.0% Black and White Biracial, 2.1% South Asian, 6.3% East Asian and 2.7% a different identity). A sensitivity power analysis conducted in the InteractionPoweR Shiny App (Baranger et al., 2023) indicated that a sample of N = 1079 (approximately 540 per condition) provides ~ 93% power to detect an interaction effect of r = .13, half the size of the interaction effect reported by Dunton and Fazio (1997).

[7] We first collected 1080 participants. Out of these, 20 had a subject ID error and 10 were excluded based on our preregistered exclusion criteria. Following our preregistered recruitment plan, we set out to collect data from 31 additional participants. We collected 30 participants, which due to an error all had duplicate responses; we redid the collection and two had incomplete data.

Procedure and Materials

As in Experiment 3, participants were randomly assigned to one of two between-subjects conditions in which they completed either the IAT or an explicit exemplar measure. All participants additionally completed the explicit category measure in counterbalanced order, and finally the MCPRS, which always came last. The explicit category measure and explicit exemplar measure were identical to Experiment 2.The IAT was identical to Experiment 2, the only difference being that we reduced the number of trials in each of the combined blocks from 80 to 40.[8] At the end of the study, participants completed an attention check and demographic questions.

Results

Descriptive statistics of and zero-order correlations between measures are presented in Table 10. Within each condition, the racial-bias measures were posistively correlated with the other racial-bias measures. MCPRS scores showed significant negative correlations with the explicit measures of racial bias, but not with the IAT.

Following our preregistered analysis plan, we conducted separate multiple-regression analyses for each condition.[9] In the IAT condition, explicit category evaluations were regressed onto standardized MCPRS scores, standardized IAT scores, and the interaction term. The main effect of MCPRS was significant, $\beta = -0.31$, $t(536) = -7.12$, $p < .001$, and the main effect of IAT scores was significant, $\beta = 0.25$, $t(536) = 5.86$, $p < .001$. Critically, the interaction term was not significant, $\beta = -0.004$, $t(540) = -0.09$, $p = .928$, indicating a positive association between the

[8] The reason for this decision was that in Experiment 2 we had to break down the task into parcels to run the SEM. In Experiment 4, there was no need for parcels, so we did not need as many trials.

[9] Two additional preregistered analyses on the factorial structure of the MCPRS are reported in the Supplemental Materials.

IAT scores and the explicit category measure regardless of motivation to control prejudiced reactions (see Figure 7).

In the explicit-exemplar condition, explicit category evaluations were regressed onto standardized MCPRS scores, the standardized explicit-exemplar evaluations, and the interaction term. The main effect of MCPRS was not significant, $\beta = -0.02$, $t(535) = -0.67$, $p = .502$, and the main effect of explicit exemplar evaluations was significant, $\beta = 0.70$, $t(535) = 18.93$, $p < .001$. Critically, the interaction term was not significant, $\beta = -0.05$, $t(535) = -1.59$, $p = .112$, indicating a positive association between explicit exemplar evaluations and the explicit category measure, regardless of motivation to control prejudiced reactions (see Figure 8).[10]

Discussion

Different from the findings with the AMP in Experiment 3, the current study found no significant effect of motivation to control prejudiced reactions on the association between IAT scores and explicit category evaluations. Instead, the IAT behaved more like an explicit exemplar measure, which also showed no interaction with motivation prejudiced reactions in the prediction of explicit category evaluations. Thus, together with the findings of the preceding studies, the results of Experiment 4 corroborate our conclusion that the exemplar-category distinction does indeed matter for the IAT, challenging the dominant narrative in the literature that the category labels in the IAT make it a category measure (De Houwer, 2001; Fazio & Olson, 2003).

[10] Although not preregistered, we also regressed explicit category evaluations onto standardized MCPRS scores, standardized exemplar evaluations, dummy-coded type of exemplar measures (i.e., implicit = 1 vs. explicit = 0), and all interactions between the three predictors. The three-way interaction was not statistically significant, $\beta = .04$, $t(1071) = 0.81$, $p = .419$, suggesting that the moderating effect of MCPRS did not depend on the type of exemplar measure.

General Discussion

Implicit measures often show dissociations from explicit self-report measures, including low correlations between the two kinds of measures (Nosek, 2005), distinct antecedents (Gawronski & Bodenhausen, 2006), and distinct behavioral correlates (Friese et al., 2008). These dissociations are commonly attributed to differences in automaticity features, in that explicit measures capture intentional responses that are relatively easy to control, whereas implicit measures capture unintentional responses that are more difficult to control. However, interpretations of the observed dissociations in terms of automaticity features presuppose that measurement type is not confounded with other important differences. Expanding on concerns that explicit measures of racial bias tend to focus on abstract social categories whereas many implicit measures of racial bias involve presentations of specific exemplars (Gawronski, 2019), the current work investigated whether dissociations between implicit and explicit measures of racial bias are at least partly accounted for by their differential focus on exemplars versus categories. To this end, we examined associations between AMP and IAT measures of racial bias with explicit measures of category and exemplar evaluations, respectively (Experiments 1 and 2). In addition, we examined for the AMP and the IAT whether motivation to control prejudiced reactions moderates associations between exemplar-focused and category-focused measures of racial bias regardless of whether the exemplar-focused measure is implicit or explicit (Experiments 3 and 4).

Our findings indicate that, although methodologically important, the distinction between exemplars and categories does not seem to matter much for dissociations between the AMP and explicit measures of racial bias. Although correlations between AMP scores and explicit measures of racial bias tended to be slightly larger when the two measures corresponded in terms of their content (i.e., both exemplar or both category) than when they did not correspond (i.e., one exemplar and one category), the obtained differences were very small overall and statistically significant in only one of the four cases (Experiment 1). Moreover, motivation to control prejudiced reactions moderated the association between an exemplar-focused AMP measure and explicit category evaluations, but motivation to control prejudiced reactions did not show the same effect on the association between explicit exemplar and explicit category evaluations (Experiment 2). Together, these findings suggest that, although AMP measures of racial bias typically include only exemplars and no references to social categories, the distinction between exemplars and categories plays a negligible role for dissociations with explicit category measures. Thus, prior evidence for dissociations between the AMP and explicit category measures may indeed be rooted in the different types of measures (i.e., implicit vs. explicit) rather than the measured contents (i.e., exemplar vs. category).

Our findings suggest a remarkably different conclusion for the IAT. Counter to the widespread assumption that the category labels in the IAT make it a measure of category evaluations (e.g., De Houwer, 2001; Fazio & Olson, 2003), IAT scores of racial bias showed a significant positive association with explicit exemplar evaluations after controlling for explicit category evaluations, but IAT scores of racial bias were unrelated to explicit category evaluations after controlling for explicit exemplar evaluations (Experiment 3). Moreover, different from the results obtained with the AMP, motivation to control prejudiced reactions did not moderate the association between IAT scores and explicit category evaluations. Instead, racial bias on the IAT showed a significant positive association with explicit category evaluations regardless of motivation to control prejudiced reactions, similar to our findings for explicit exemplar evaluations. Together, these results suggest that the distinction between categories and exemplars does matter for the IAT, and that the IAT is functionally closer to measures of exemplar evaluations than measures of category evaluations. Thus, different from our conclusion for the AMP, IAT measures of racial bias may show dissociations with explicit category measures because of the exemplar content of the IAT rather than the implicit versus explicit nature of the measurement instruments.

Relation to Prior Findings

The current findings conflict with a dominant narrative in the literature on implicit measures, suggesting that the category labels in the IAT make it a measure of category evaluations (e.g., De Houwer, 2001; Fazio & Olson, 2003). In contrast, sequential priming tasks such as the AMP and the EPT are assumed to capture exemplar evaluations, in that the standard variants of the these tasks involve presentations of specific exemplars without any reference to social categories (e.g., Livingston & Brewer, 2002; Olson & Fazio, 2003). To reconcile the conflict between these assumptions and the current findings, it is worth taking a closer look at two seminal studies that are frequently cited in support of the dominant narrative.

One study by De Houwer (2001) suggests that the valence of specific exemplars used as target stimuli in the IAT does not influence responses on the IAT. In this study, British participants completed an IAT to measure preference for British versus foreign people, with names of liked British people (e.g., Princess Diana), disliked British people (e.g., Margaret Thatcher), liked non-British people (e.g., Mahatma Ghandi), and disliked non-British people (e.g., Adolf Hitler) as target stimuli. Likeability and nationality of the targets were manipulated within-subjects, such that all participants were presented with an equal number of names for each of the four target groups. Participants were asked to classify the targets as British versus foreign in the IAT. The central finding was that participants responded faster when British names shared the same response key with positive words and foreign names shared the same response key with negative words, compared with the reversed mapping. Importantly, the observed difference in response times was not significantly affected by the valence of the targets. These findings are commonly interpreted as evidence that the IAT measures evaluations of the categories in the task rather the specific exemplars used as target stimuli.

To reconcile the apparent conflict with the current findings, it is worth considering several important aspects of De Houwer's (2001) study. First, the main conclusion in De Houwer's study is based on a null effect obtained with a rather small sample of 28 participants. Although the within-subjects manipulation of target valence is advantageous for statistical power, the sample would require an effect size of *dz* = .55 for the detection of a significant difference with 80% power in a two-tailed test. If one aims for 95% power to bolster a meaningful interpretation of a null effect, the effect size would have to be *dz* = .71. Thus, whether the study had sufficient statistical power for conclusions based on a null effect seems debatable. Second, even if the study had sufficient statistical power for a meaningful interpretation of a null effect, the design of De Houwer's study speaks only to the question of whether characteristics of individual exemplars influence responses in the IAT. It does not speak to the current question of whether aggregated responses across multiple exemplars in the IAT show greater correspondence to explicit measures of aggegrated responses to the same exemplars or explicit measures of abstract category evaluations. From this perspective, the null effect in De Houwer's study (assuming it is meaningful despite the small sample) does not conflict with the current conclusion that IAT scores of racial bias show greater correspondence with explicit measures of aggregated responses to the same exemplars than explicit measures of abstract category evaluations.

Another important study that is frequently cited in support of the dominant narrative (Olson & Fazio, 2003) found that EPT and IAT measures of racial bias showed a stronger association when participants were asked to keep a mental tally of how many Black and White faces are presented in the EPT than when they completed the EPT without instructions to focus on social categories. According to the authors, these findings suggest that low correlations between the IAT and EPT are due to differing attitude objects in the two tasks, in that the IAT measures responses to abstract catgeories whereas the EPT measures responses to specific exemplars of a given category (see also Fazio & Olson, 2003). Yet, if participants are instructed to think of the exemplars presented in the EPT in terms of abstract social categories (as is the case in the IAT), the EPT becomes sensitive to category evaluations, which increases its correspondence with the IAT. Although this interpretation seems intuitively plausible, a rarely acknowledged aspect of Olson and Fazio's (2003) findings is that the internal consistency of the EPT was modest when participants were asked to focus on abstract catgeories (split-half correlation of *r* = .39), yet close to zero when they were not asked to focus on abstract categories (split-half correlation of *r* = .04). While low internal consistency does not necessarily undermine a measure's sensitivity to experimental effects, it is known to suppress correlations with other measures, including measures of the same construct (see Koppehele et al., 2020). Thus, a simple alternative interpretation of Olson and Fazio's finding is that EPT scores obtained under standard instructions were too noisy to show a meaningful association with the IAT, and that instructions to focus on the category membership of the exemplars in the EPT increased its association with the IAT by increasing the reliability of the EPT. Consistent with this interpretation, Gawronski et al. (2010) found that the EPT provided (1) reliable scores of race bias only when participants focused on the race of the presented exemplars but not when they focused on the age of the presented exemplars and (2) reliable scores of age bias only when participants focused on the age of the presented exemplars but not when they focused on the race of the presented exemplars. Importantly, the AMP provided reliable scores of both race and age bias regardless of whether participants focused on the race or age of the presented exemplars. Together, these results suggest that Olson and Fazio's (2003) findings may be due to the low reliability of the EPT under standard task instructions, rather than a unique sensitivity of the IAT to evaluations at the category level. From this perspective, Olson and Fazio's (2003) findings do not contradict the current finding that IAT scores of racial bias show greater correspondence with explicit measures of aggregated responses to the same exemplars than explicit measures of abstract category evaluations. They also do not contradict the current findings that the exemplar-category distinction does not seem to matter for the AMP.

Another important question is how the current findings can be reconciled with earlier findings showing that associations between implicit and explicit measures of bias tend to be stronger for participants with a weak motivation to control prejudiced reactions compared to participants with a strong motivation to control prejudiced reactions (e.g., Akrami & Ekehammar, 2005; Degner & Wentura, 2008; Dunton & Fazio, 1997; Fazio et al., 1995; Gawronski et al., 2003; Payne et al., 2005). In the current work, we replicated this widely cited pattern for the AMP but not for the IAT. Regarding the null effect obtained for the IAT, it is worth noting that prior studies on the presumed effect of motivation to control prejudiced reactions used sample sizes that, by today's standards, may be deemed insufficient for the detection of the hypothesized two-way interaction (see e.g., da Silva Frost & Ledgerwood, 2020) with samples sizes ranging between *N* = 42 (Akrami & Ekehammar, 2005) and *N* = 111 (Fazio et al., 1995). We are aware of five published studies that found a significant interaction between

IAT-measured evaluative bias and MCPRS scores in the prediction of explicit category evaluations, and all these studies seem underpowered for the detection of the predicted two-way interaction with sample sizes of *N* = 42 (Akrami & Ekehammar, 2005), *N* = 69 (Gawronski et al., 2003), *N* = 87 (Hofmann, Gschwendner, et al., 2005, Study 2), *N* = 93 (Hofmann, Gschwendner, et al., 2005, Study 1), and *N* = 103 (Ziegert & Hanges, 2005), respectively. Although we cannot rule that the non-significant interaction between MCPRS scores and IAT-measured racial bias in Experiment 4 reflects a false negative, the large sample size in that study (*N* = 1079) renders such an interpretation unlikely. Instead, it seems more likely that prior findings suggesting an interaction between IAT-measured evaluative bias and MCPRS scores in the prediction of explicit category evaluations are false positives (see Button et al., 2013). Indeed, the non-significant interaction in Experiment 4 is perfectly consistent with our findings that (1) IAT scores of racial reflect evaluations at the level of exemplars rather than categories (Experiment 2), and (2) evaluations at the exemplar-level do not show the same interaction effect with motivation to control prejudiced reactions (Experiments 3 and 4).

Implications

The current findings have important implications for racial-bias research comparing responses on implicit and explicit measures. Dissociations between the two kinds of measures are typically interpreted in terms of automaticity features, in that implicit measures are assumed to capture unintentional responses that are relatively difficult to control whereas explicit measures capture intentional responses that are relatively easy to control. However, such interpretations require that implicit and explicit measures do not differ in terms of other features, such as their specific content. While this issue has been long acknowledged in research using implicit and explicit measures to study personality self-concepts (e.g., Asendorpf et al., 2002;

Back et al., 2009; Peters & Gawronski, 2011), it has been largely ignored in research on racial bias, where implicit measures commonly include presentations of specific exemplars while explicit measures predominantly focus on abstract social categories without presentations of specific exemplars (for discussions, see Gawronski, 2019; Payne et al., 2008). The current findings suggest that, while the distinction between categories and exemplars does not matter much for the AMP, it does seem essential for the IAT, in that the IAT measures evaluative responses at the level of exemplars rather than abstract social categories. Hence, prior findings showing low correlations between IAT measures of social bias and explicit self-report measures (for a meta-analysis, see Hofmann, Gawronski, et al., 2005), distinct antecedents (for a meta-analysis, see Forscher et al., 2019), and distinct behavioral correlates (for a meta-analysis, Kurdi et al., 2019) may have nothing to do with commonly invoked difference between types of measures (i.e., implicit vs. explicit). Instead, the same dissociations may emerge for explicit measures of exemplar and category evaluations, which would suggest fundamentally different conclusions for both theory and practice. Based on these considerations, we recommend that future research using implicit and explicit measures of social bias should always include an explicit exemplar measure in addition to the commonly used explicit category measures. Such designs permit stronger conclusions about whether dissociations are driven by the different types of measures or differences between responses to categories and exemplars.

Constraints on Generality

Despite several important strengths (e.g., carefully controlled experimental setting, preregistration, large sample sizes), the current work also has some notable limitations. First, the current work focused exclusively on the AMP and the IAT, which are only two instruments among the large set of the currently available implicit measures (for revies, see Gawronski & De

Houwer, 2014; Greenwald & Lai, 2020). In addition to their greater prominence, our focus on the AMP and the IAT was based on the facts that (1) high internal consistency is methodologically imperative for the correlational designs of the current studies (see Koppehele-Gossel et al., 2020) and (2) the AMP and the IAT are the only two implicit measures that meet this criterion. Nevertheless, future research investgating the confound between type of measure and type of content for other implicit measures would be extremely valuable.

Second, the current work focused exclusively on measures of racial bias involving evaluative responses to Black and White targets. Yet, social biases exist for a broad range of groups, and not all of these biases involve evaluative responses (Amodio & Devine, 2006). Because the confound between type of measures and type of content seems relevant for all of these cases, future studies on social biases against other groups involving non-evaluative dimensions (e.g., semantic gender stereotypes) would be helpful to address the generality of our findings.

Finally, all of the reported studies have been conducted with participants from the United States, which raises the question of whether the obtained results would replicate in samples from other countries. For example, some researchers have argued that racial categories tend to be more salient in the United States compared to many European countries (e.g., Degner & Wentura, 2010), which may affect the interplay of racial bias at the exemplar-level and the category-level in either explicit or implicit measures (or both). Thus, future research investigating the reproducibility of our findings with samples of non-American participants would be helpful to gauge the generalizability of our conclusions.

Conclusion

The current research addressed a common confound in research on racial bias: the confound between type of measure (i.e., implicit vs. explicit) and type of content (i.e., exemplar vs. category). Our findings suggest that, while the exemplar-category distinction does not seem to matter much for the AMP, it does matter for the IAT in that the IAT behaves more like a measure of exemplar evaluations than category evaluations. These findings raise important questions about whether previously obtained dissociation between self-report and IAT measures of racial bias are driven by the explicit versus implicit nature of the instruments or their different contents. Based on our conclusions, we recommend that future research comparing racial bias on implicit and explicit measures include explicit measures at both the category- and the exemplar-level.

References

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

Ajzen, I., and 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: Erlbaum.

Akrami, N., & Ekehammar, B. (2005). The association between implicit and explicit prejudice: The moderating role of motivation to control prejudiced reactions. *Scandinavian Journal of Psychology*, *46*, 361-366.

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93*, 203-231.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*, 652–661.

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology, 83*, 380-393.

Axt, J. R. (2018). The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science*, *9*, 896-906.

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology, 97*, 533-548.

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General, 147*, 1264–1272.

Baranger, D. A., Finsaas, M., Goldstein, B., Vize, C., Lynam, D., & Olino, T. M. (2023). Tutorial: Power analyses for interaction effects in cross-sectional regressions. *Advances in Methods and Practices in Psychological Science, 6,* 1-13.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 1-40). Hillsdale, NJ: Erlbaum.

Brooks, L. R. (1987). Decentralized control of cognition: The role of prior processing episodes. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 141-174). Cambridge, England: Cambridge University Press.

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E., & Munafo, M. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Review Neuroscience, 14*, 365–376

Cameron, C. D., Brown-Iannuzzi, J., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behaviors and explicit attitudes. *Personality and Social Psychology Review, 16,* 330-350.

Cooley, E., & Payne, B. K. (2019). A group is more than the average of its parts: Why existing stereotypes are applied more to the same individuals when viewed in groups than when viewed alone. *Group Processes & Intergroup Relations, 22*, 673-687.

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review*, *24*, 212-232.

da Silva Frost, A., & Ledgerwood, A. (2020). Calibrate your confidence in research findings: A tutorial on improving research methods and practices. *Journal of Pacific Rim Psychology*.

da Silva Frost, A., Wang, Y. A., Eastwick, P. W., & Ledgerwood, A. (2022). Summarized attribute preferences have unique antecedents and consequences. *Journal of Experimental Psychology: General.* Advance online publication. https://doi.org/10.1037/xge0001242

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*, 443-451.

De Houwer, J., & Boddez, Y. (2022). Bias in implicit measures as instances of biased behavior under suboptimal conditions in the laboratory. *Psychological Inquiry*, *33*, 173-176.

Degner, J., & Wentura, D. (2008). The extrinsic affective Simon task as an instrument for indirect assessment of prejudice. *European Journal of Social Psychology, 38,* 1033-1043.

Degner, J., & Wentura, D. (2010). Automatic prejudice in childhood and early adolescence. *Journal of Personality and Social Psychology, 98*, 356-374.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23,* 316-326.

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*, Article 103837.

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.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition & Emotion*, *15*, 115-141.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*, 1013–1027.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297-327.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*, 522-559.

Friese, M., Hofmann, W., & Schmitt, M. (2008). When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology*, *19*, 285-338.

Fujita, K., & Han, H. A. (2009). Moving beyond deliberative control of impulses: The effect of construal levels on evaluative associations in self-control conflicts. *Psychological Science, 20,* 799-804.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14,* 574-595.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692-731.

Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorization influence spontaneous evaluations of multiply categorizable objects? *Cognition and Emotion, 24,* 1008-1025.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd edition, pp. 283-310). New York: Cambridge University Press.

Gawronski, B., De Houwer, J., & Sherman, J. W. (2020). Twenty-five years of research using implicit measures. *Social Cognition, 38,* s1-s25.

Gawronski, B., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment*, *24,* 218-225.

Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology*, *33*, 573-589.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022a). Implicit bias ≠ bias on implicit measures. *Psychological Inquiry*, *33*, 139-155.

Gawronski, B., Ledgerwood, A., & Eastwick, P. W. (2022b). Reflections on the difference between implicit bias and bias on implicit measures. *Psychological Inquiry*, *33*, 219-231.

Gawronski, B., & Ye, Y. (2014). What drives priming effects in the affect misattribution procedure? *Personality and Social Psychology Bulletin*, *40*, 3-15.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4-27.

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, *71*, 419-445.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197-216.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*, 1369–1392.

Hansen, J., & Trope, Y. (2013). When time flies: How abstract and concrete mental construal affect the perception of time. *Journal of Experimental Psychology: General, 142*, 336-347.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and social psychology bulletin*, *31*, 1369-1385.

Hofmann, W., Gschwendner, T., & Schmitt, M. (2005). On implicit–explicit consistency: The moderating role of individual differences in awareness and adjustment. *European Journal of Personality*, *19*, 25-49.

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (2020). Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology, 87,* Article 103905.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist, 74*, 569-586.

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.

Ledgerwood, A., Trope, Y., & Chaiken, S. (2010). Flexibility now, consistency later: Psychological distance and construal shape evaluative responding. *Journal of Personality and Social Psychology, 99*, 32-51.

Ledgerwood, A., & Wang, Y. A. (2018). Achieving local and global shared realities: Distance guides alignment to specific or general social influences. *Current Opinion in Psychology*, *23*, 62-65.

Levesque, H. J., Davis, E., & Morgenstern, L. (2012, June). The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning* (pp. 552-561).

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category- based processing of facial stimuli. *Journal of Personality and Social Psychology, 82*, 5-18.

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.

McCrea, S. M., Wieber, F., & Myers, A. (2012). Construal level mind-sets moderate self- and social stereotyping. *Journal of Personality and Social Psychology, 102*, 51–68.

Melnikoff, D. E., & Kurdi, B. (2022). What implicit measures of bias can do. *Psychological Inquiry*, *33*, 185-192.

Moors, A., & De Houwer, J. (2006). Automaticity: A conceptual and theoretical analysis. *Psychological Bulletin, 132*, 297-326.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*, 565-584.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences*, *15*, 152-159.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test. *Experimental Psychology*, *54*, 14-29.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14*, 636-639.

Olson, M. A., & Gill, L. J. (2022). Commentary on Gawronski, Ledgerwood, and Eastwick, Implicit Bias ≠ Bias on Implicit Measures. *Psychological Inquiry*, *33*, 199-202.

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94*, 16-31.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277-293.

Payne, B. K., & Lundberg, K. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass, 8*, 672-686.

Peters, K. R., & Gawronski, B. (2011). Mutual influences between the implicit and explicit self-concepts: The role of memory activation and motivated reasoning. *Journal of Experimental Social Psychology*, *47*, 436-442.

Ratliff, K. A., & Smith, C. T. (2022). Implicit bias as automatic behavior. *Psychological Inquiry*, *33*, 213-218.

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.

Smith, E. R. (1998). Mental representation and memory. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology*, Vol. 1 (4th ed., pp. 391–445). New York: Oxford University Press.

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99*, 3-21.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*, 440-463.

Wakslak, C., & Trope, Y. (2009). The effect of construal level on subjective probability estimates. *Psychological Science, 20*, 52-58.

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.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology, 72*, 262–274.

Ziegert, J. C., & Hanges, P. J. (2005). Employment Discrimination: The Role of Implicit Attitudes, Motivation, and a Climate for Racial Bias. *Journal of Applied Psychology, 90*, 553–562.

Table 6. Descriptive statistics of racial-bias scores as a function of measurement type (explicit vs. implicit) and measurement content (exemplar vs. category), Experiment 1.

Measure	*n*	*M*	*SD*	*α*
Explicit Category	611	0.21	1.22	.76
Explicit Exemplar	510	-0.01	1.01	.91
Implicit Category	564	0.02	0.29	.72
Implicit Exemplar	557	0.02	0.27	.65

Table 7. Correlations between measures of racial bias as a function of measurement type (explicit vs. implicit) and measurement content (exemplar vs. category), Experiment 1.

Condition	*n*	*r*	95% CI	*p*
A. Explicit Exemplar - Implicit Exemplar	264	.50	[.40, .58]	< .001
B. Explicit Exemplar - Implicit Category	246	.41	[.30, .51]	< .001
C. Explicit Category - Implicit Category	318	.56	[.48, .63]	< .001
D. Explicit Category - Implicit Exemplar	293	.45	[.35, .54]	< .001

Table 8. Descriptive statistics of and zero-order correlations between measures, Experiment 2.

Measure	*n*	*M*	*SD*	α	1	2	3
1. Explicit Category	585	-0.50	1.16	.94	-		
2. Explicit Exemplar	585	-0.35	1.02	.87	.71***	-	
3. IAT	585	0.33	0.42	.95	.29***	.33***	-

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

Table 9.

Descriptive statistics of and zero-order correlations between measures as a function of measurement condition (explicit exemplar vs. AMP), Experiment 3.

Measure	*M*	*SD*	*α*	1	2	3
Explicit-Exemplar (*n* = 586)						
Explicit Category	0.03	1.10	.68	-		
MCPRS	0.50	0.80	.87	-.15***	-	
Explicit Exemplar	-0.22	0.88	.92	.62***	-.15***	-
AMP (*n* = 559)						
Explicit Category	0.06	1.13	.64	-		
MCPRS	0.56	0.83	.88	-.11*	-	
AMP	0.16	0.26	.72	.33***	-.11**	-

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

Table 10.

Descriptive statistics of and zero-order correlations between measures as a function of measurement condition (explicit exemplar vs. IAT), Experiment 4.

Measure	*M*	*SD*	*α*	1	2	3
Explicit-Exemplar (*n* = 539)						
1. Explicit Category	-0.47	1.10	.93	-		
2. MCPRS	0.54	0.90	.90	-.14***	-	
3. Explicit Exemplar	-0.24	0.84	.90	.65***	-.19***	-
IAT (*n* = 540)						
1. Explicit Category	-0.42	1.07	.95	-		
2. MCPRS	0.55	0.83	.87	-.29***	-	
3. IAT	0.38	0.39	.66	.24***	-.005	-

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

Figure 3.

Illustration of potential associations between measures of racial bias as a function of measurement type (implicit vs. explicit) and measurement content (exemplar vs. category).

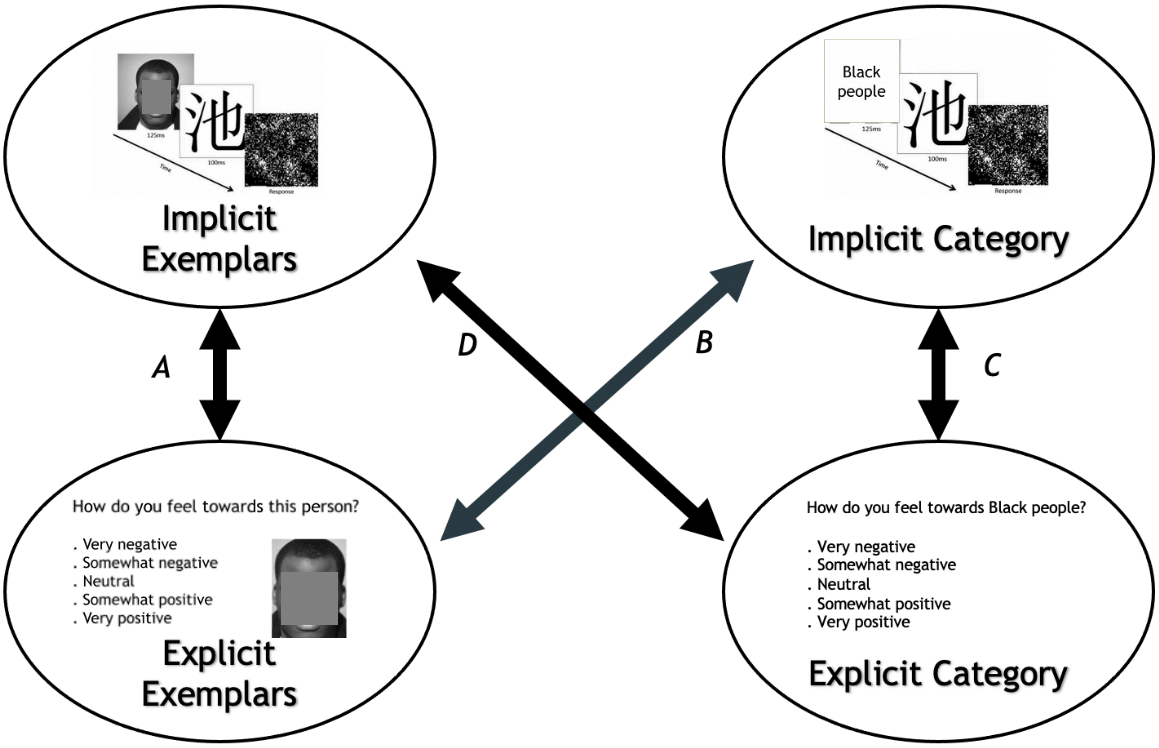


Figure 4.

Results of structural-equation modeling predicting racial bias on the IAT via explicit exemplar evaluations and explicit category evaluations.

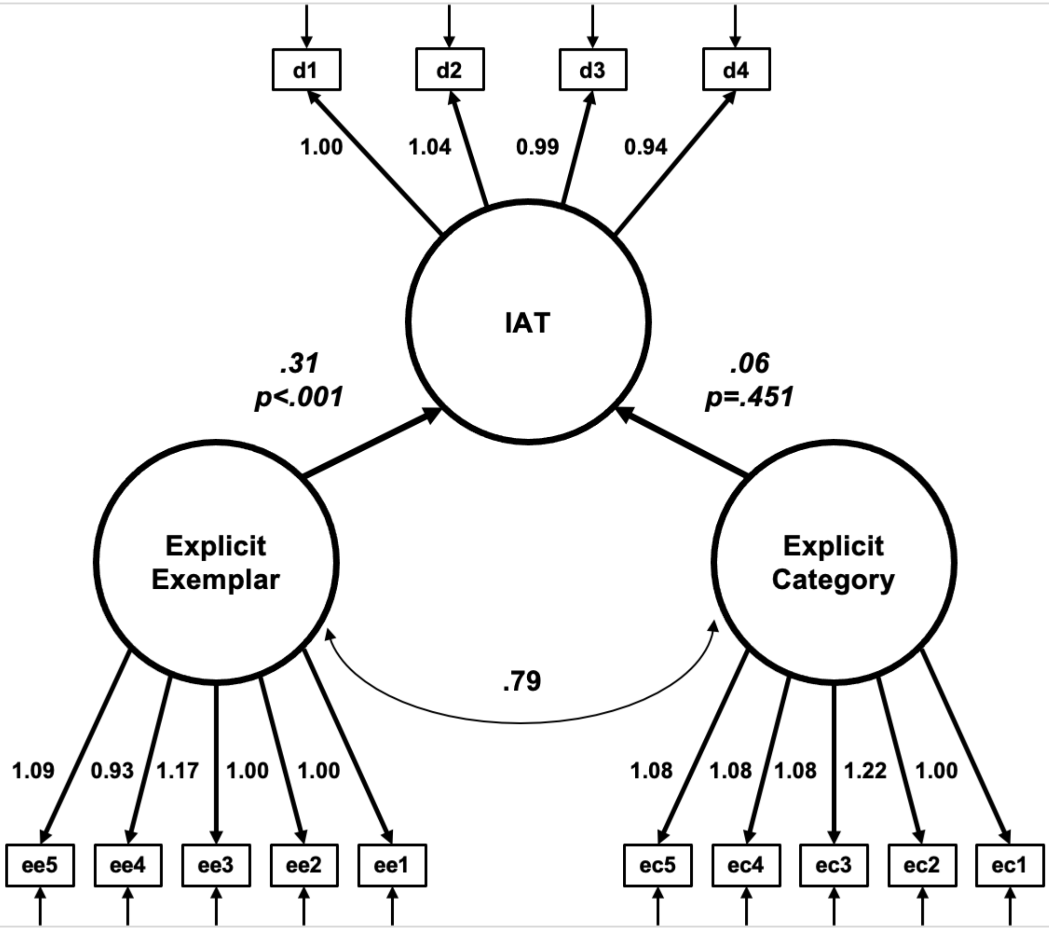


Figure 5.

Explicit category evaluations as a function of racial bias on the AMP and motivation to control prejudiced reactions (MCPRS), Experiment 3.

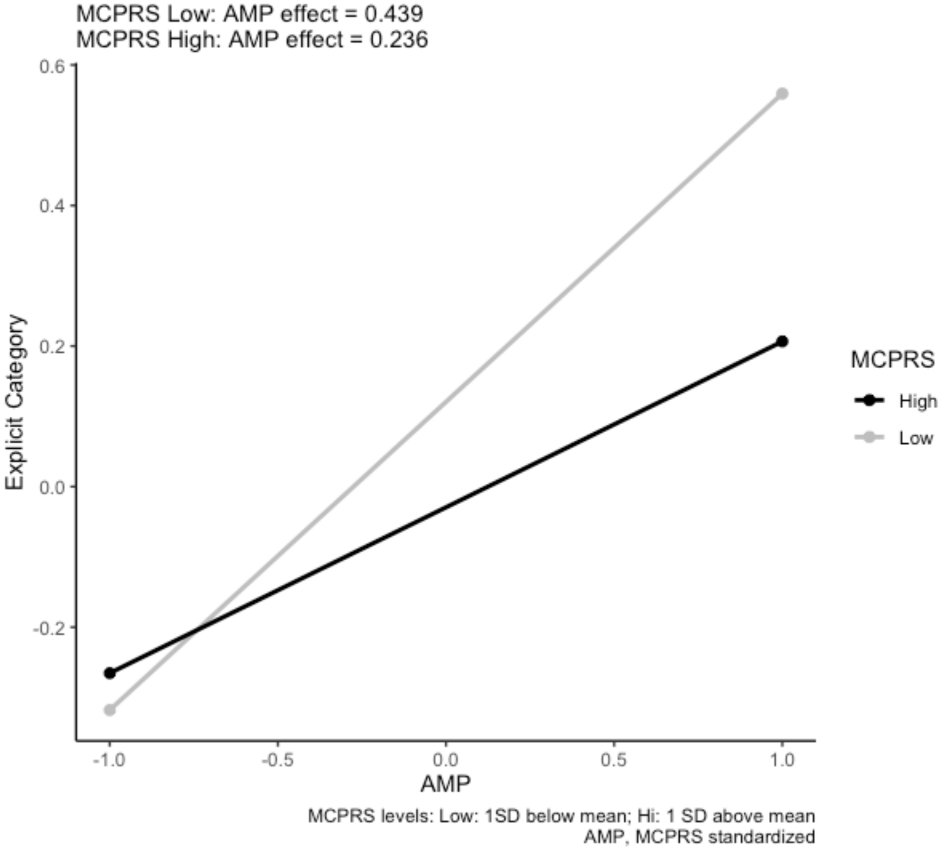


Figure 6.

Explicit category evaluations as a function of explicit exemplar evaluations and motivation to control prejudiced reactions (MCPRS), Experiment 3.

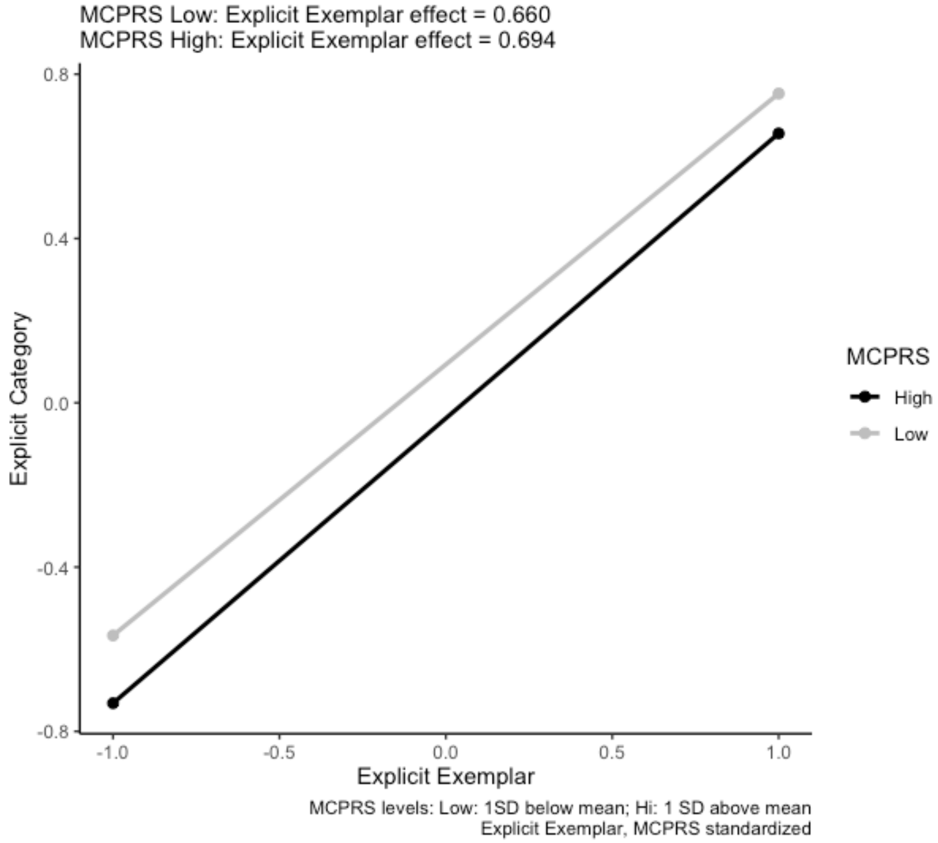


Figure 7.

Explicit category evaluations as a function of racial bias on the IAT and motivation to control prejudiced reactions (MCPRS), Experiment 4.

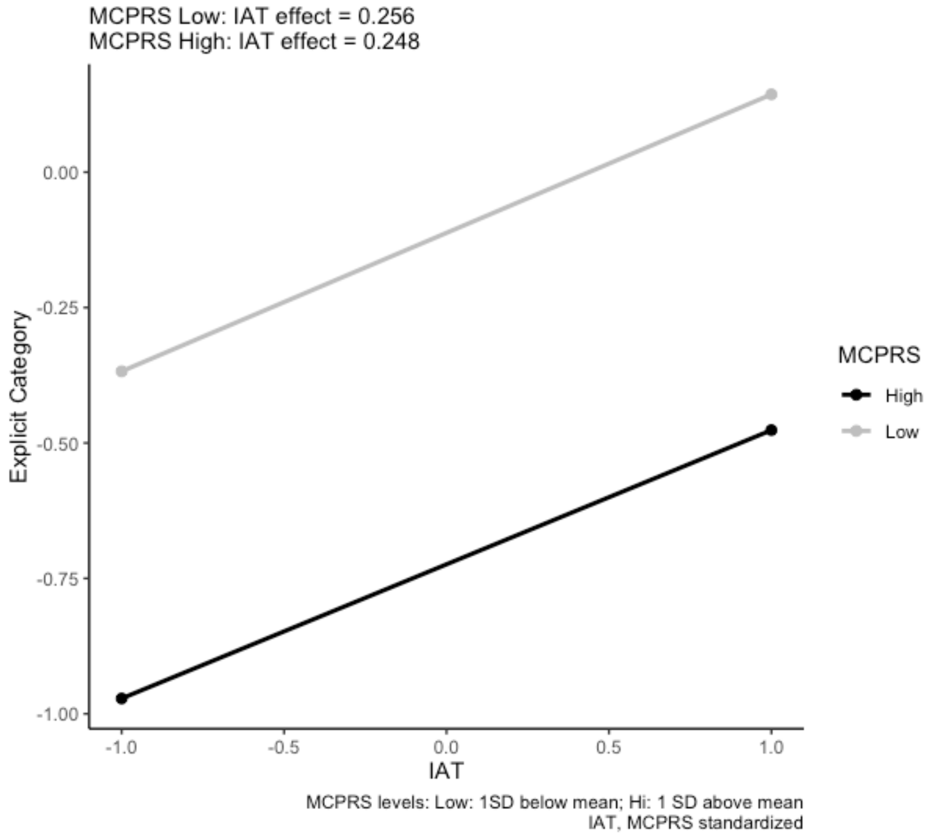
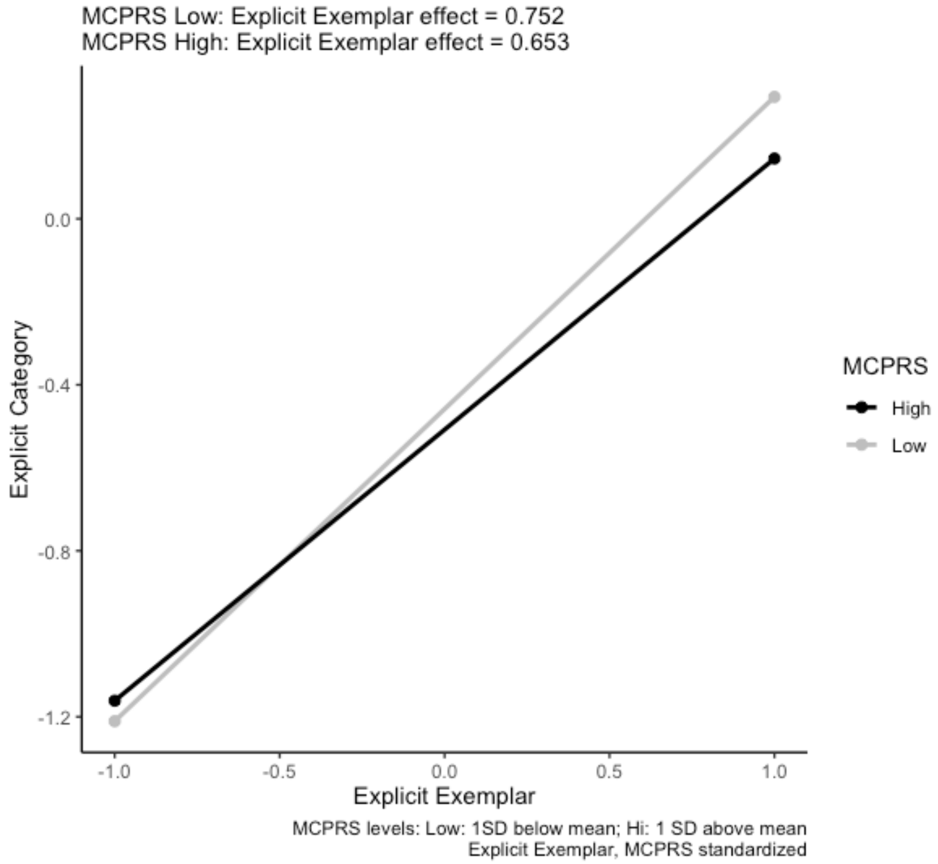


Figure 8.

Explicit category evaluations as a function of explicit exemplar evaluations and motivation to control prejudiced reactions (MCPRS), Experiment 4.



Chapter 3

Experimental Tests of the Role of Ideal Partner Preferences in Relationships

Abstract

A large literature on ideal-partner preferences (i.e., the extent to which people positively evaluate a trait in a romantic partner) suggests that ideals are important in people's romantic lives. Nevertheless, the consequences of ideals have rarely been tested experimentally. In this research, we (a) successfully developed a paradigm to reliably manipulate preferences for traits in a romantic partner, and (b) conducted the first experimental tests of four key theoretical perspectives. These perspectives were: *motivated projection* (i.e., high ideals cause people to believe that their partner possesses the relevant trait); *preference-matching (trait-weighting)* (i.e., high ideals cause people to be more satisfied with a partner to the extent that they think their partner possesses the trait); *situation selection* (i.e., high ideals cause people to seek out situations where they are more likely to meet partners who possess the trait); and *perceiver effects* (i.e., high ideals cause people to believe everyone—even strangers—possess that trait). We found evidence for all four accounts to varying degrees: The motivated-projection and situation-selection accounts received very strong support, the perceiver-effects account received some support, and the preference-matching (trait-weighting) account received support in one of three studies. We were unable to detect a direct effect of the manipulation on relationship satisfaction, but perceptions of the partner's attributes did mediate the effect of the manipulation on satisfaction. Implications for our understanding of positive illusions and motivational processes in romantic relationships are discussed.

Keywords: romantic relationships; ideals; attribute preferences; positive illusions; person perception

Measuring the Impact of Multiple Social Cues to Advance Theory in Person Perception Research

Experimental Tests of the Role of Ideal Partner Preferences in Relationships

People often have elaborate ideas about what the best version of their love life should look like (Fletcher et al., 1999; Hassebrauck, 1997; Rusbult et al., 1993). One relevant scientific literature has examined ideals for a romantic partner's traits (i.e., *ideal partner preferences* or *mate preferences*). Throughout this literature, ideals are conceptualized and measured as evaluations of dimensional attributes: ratings of how positively or negatively people feel about traits like "intelligent," "kind," or "attractive" in a romantic partner (Ledgerwood et al., 2018). The consensus is that ideals play a key role in causing a variety of downstream relational consequences (Conroy-Beam & Buss, 2016; Fletcher & Simpson, 2000; Fetcher et al., 1999; Simpson, et al., 2001; Gerlach et al., 2019; LaPrelle et al., 1990; Murray et al., 1996a, 1996b).

Nevertheless, *experimental* evidence for the role of ideal partner preferences remains nearly nonexistent. We are aware of only five prior attempts to manipulate ideal partner preferences (da Silva Frost et al., 2022; Eastwick, Smith, et al., 2019; Eagly et al., 2009; Kille et al., 2013; Nelson & Morrison, 2005), and none were designed to capture any possible consequences of ideals. Our collective understanding of ideals and their effects has been limited by the lack of strong experimental manipulations. Recent advances in causal analyses of correlational data are commendable and impressive (e.g., Grosz et al., 2020, 2024; Pearl et al., 2016); nevertheless, experiments (when feasible) remain the preferred way to learn about causal effects due to their ability to eliminate alternative explanations (e.g., Murnane & Willet, 2011; Pearl, 2009; Shadish et al., 2002). Our sense was that an experimental approach to ideals was indeed feasible but had not been thoroughly explored.

This article describes a social-cognitive paradigm that researchers can use to manipulate participants' own ideal partner preferences. Furthermore, we use this manipulation to differentiate among existing theoretical models of ideal partner preferences (for other examples, see Brandner et al., 2020; Conroy-Beam et al., 2022) and to build a more comprehensive and robust understanding of the causal effects of ideal preferences.

Ideal Partner Preferences and their (Purported) Consequences

Theoretical perspectives on the possible causal consequences of ideals for traits abound. We focus on four of these perspectives in the current research, all of which are foundational in the field and underscore how ideals for traits could cause important outcomes in romantic contexts (Eastwick, Finkel, et al., 2019; Fletcher et al., 2020).

Theoretical Account #1: Motivated Projection

The first perspective, *motivated projection*, is associated with Murray and colleagues' foundational work on idealization in romantic relationships (Murray et al., 1996a, 1996b). This approach stems from a broader tradition in psychology on the causes and consequences of positive illusions. According to this literature, holding idealized views of the self, the world, and other people can lead to various positive outcomes (Taylor & Brown, 1988).

Murray and colleagues applied this logic to the romantic domain, positing that people should be motivated to project their own ideals into their partners—an illustration of "rose-colored glasses." In other words, people may be motivated to believe that their partners possess the traits that they themselves value (Murray et. al., 1996a, 1996b).

In this account, ideals can even have downstream consequences for relationship satisfaction. That is, ideals may serve as the upstream predictor in a mediational model that cascades from ideals to the positive perception of a partner's traits, to relationship satisfaction, to relationship stability. Although Murray and colleagues (1996a, 1996b) found support for these pathways longitudinally—suggesting a causal pathway—experimental demonstrations would be useful to make this case more conclusively.

Theoretical Account #2: Ideal Partner Preference-Matching (Trait Weighting)

This second perspective is central to the ideal standards model (Fletcher et al., 1999, 2020; Fletcher & Simpson, 2000) and evolutionary models which suggest that there are individual differences in the way that people weight the importance of attributes in a mate (Buss, 1989). The ideal standards model draws from the interdependence theory postulate that relationship quality is a function of discrepancies between (a) the rewards people receive from the relationship and (b) their ideals and expectations for what those rewards should be (Thibaut & Kelley, 1959). In other words, the extent to which a person's ideals matches their perception of a partner's traits should cause them to positively evaluate that partner.

To illustrate, imagine that people vary in the extent to which their ideal romantic partner is attractive: Some people say this trait is very important to them, others say it is not terribly important (e.g., Buss, 1989). The ideal standards model predicts that perceiving a romantic partner to be attractive should have stronger evaluative consequences (i.e., it should receive a stronger weighting) for people with high rather than low ideals for attractiveness. That is, the ideal × trait perception interaction should predict outcomes like attraction, relationship satisfaction, or breakup (Eastwick et al., 2014).

If ideals *cause* the consequences posited by the ideal standards model, then a manipulation of ideals should interact with a partner-trait perception to predict outcomes like relationship satisfaction. If scholars could successfully implement a manipulation that increases people's ideals for a given trait, the *manipulated portion of that ideal* (i.e., due to being in the experimental condition) should interact with people's perceptions of a partner's trait to predict higher relationship satisfaction (a condition × trait perception interaction).[11]

To be clear: There are other models of the consequences of ideal partner preference-matching that are not about weighting. These alternatives include: ideals as thresholds, ideals as comparison points, and ideals as correlated patterns across many attributes (Brandner et al., 2020; Conroy-Beam et al., 2022; Eastwick et al., 2019). The experimental approaches described in this article do not address these other models: We cannot address ideals as thresholds/comparison points because our manipulation does not translate to trait "scale points," and we cannot address ideals as correlated patterns because we are only manipulating a single trait. Therefore, our study bears only on the (nevertheless foundational) weighed ideal partner preference model.

Theoretical Account #3: Situation Selection

This account resembles the ideal-partner preference matching (trait weighting) account, but the outcome is not an evaluation or selection of a specific partner. Instead, the outcome is the evaluation or selection of an environment, setting, or situation where desirable partners might be found. In other words, ideals might cause people to pursue certain ways of shaping their own "field of eligibles" (Schellenberg, 1960) so that the possible partners they encounter tend to match their ideal partner preferences on average.

A series of studies by da Silva Frost et al. (2022) found that ideals for a given attribute correlated with participants' interest in joining a website featuring potential partners with high

[11] An experimental manipulation of ideals could have two kinds of interactive consequences. First, the manipulation could operate according to the preference-matching process described here (i.e., condition × trait perception = satisfaction). Alternatively, the manipulation could operate like a moderated form of the motivated projection process (i.e., theoretical account #1) such that the experimental manipulation is especially effective for participants who are currently satisfied in their relationships (i.e., condition × satisfaction = trait perception). We will test both forms of these interactive predictions in this paper.

levels of the attribute (with *r*s ranging from .30-.50). For example, participants who had strong ideals for “intelligence” in a romantic partner were more likely to be interested in joining a website described as “providing access to partners in the top 30% of intelligence.” Suggestive evidence also comes from a study by Kurzban and Weeden (2007), who found that participants who idealized certain attributes were more likely to choose to attend speed-dating events featuring dates with that attribute, as long as those attributes were knowable ahead of time (e.g., the event was open to attendees of a certain age or ethnicity). Nevertheless, since ideals were measured variables in these studies, it’s unclear if these ideals actually caused participants to pursue this form of situation selection.

Theoretical Account #4: Perceiver Effects

A fourth perspective revolves around perceiver effects, individual differences in the way that people perceive traits in others. According to this perspective, people differ in their general trait-perceiving tendencies (e.g., they believe that others are generally attractive), and apply these tendencies whenever they meet and evaluate others (Kenny, 1994; Srivastava et al., 2010). This idea—that people have idiosyncratic views of the generalized “other”—has long been influential in psychology, stemming from Erikson’s stages of development and attachment theory (Erikson, 1959; Bowlby 1988). In person perception studies, researchers have commonly found that perceiver effects account for 20-30% of the variance in people’s Big Five trait judgments (e.g., Jamie thinks others are generally extraverted, whereas Max thinks others are generally introverted), and furthermore, these tendencies tend to be stable over time (Rau et al., 2021).

Relatedly, people’s perceptual tendencies are attuned to their attitudes: When people have strong attitudes about something, they tend to notice it more often (Roskos-Ewoldsen & Fazio, 1992). Thus, a person with a strong positive ideal for a trait may notice that trait more readily in others, increasing the likelihood they rate a romantic partner highly on that trait. An analysis of speed-dating data found that ideals correlated modestly with the extent participants viewed *all their dates* highly on a given trait, whether they liked that date or not (*r* = .16; Eastwick, Finkel, et al., 2019). This perspective's novel prediction is that ideals should generally be associated with a tendency to rate others highly on a given trait—not just romantic partners, but other people, too.

Table 11.

Four Causal Models of Ideal Partner Preferences

Theoretical Account		Description	References
1	Motivated perception	High ideals for a trait cause me to believe that my partner possesses that trait.	Murray et al. (1996a, 1996b)
2	Preference-matching (trait weighting)	High ideals cause me to be more attracted to/satisfied with my partner to the extent that I think my partner possesses that trait.	Eastwick et al., (2014); Fletcher et al. (1999)
3	Situation selection	High ideals cause me to seek out situations where I am more likely to meet partners who possess that trait.	da Silva Frost et al. (2022); Kurzban & Weeden (2007)
4	Perceiver effects	High ideals cause me to believe everyone possesses that trait.	Eastwick, Finkel, et al., (2019); Rau et al. (2021)

Note: Theoretical account #1 also notes that motivated perception effects could be stronger to the extent that the participant is more satisfied with/feels psychologically closer to the partner.

In summary, by manipulating ideal partner preferences, we can compare these four perspectives vis a vis the strength of the causal role of ideals in these processes.[12]

[12] These four accounts correspond to mechanisms #5, #7, #6, and #4, respectively, in Eastwick, Finkel, et al. (2019, Table 2). In mechanisms #1-#3 in that table, ideals are not causal.

Creating a Replicable and Robust Manipulation of Ideal Partner Preferences

One goal of this paper was to develop and replicate a manipulation of ideal partner preferences. Drawing from the social cognitive tradition, we modified a covariation detection paradigm that has successfully been used in prior research to shift attitudes toward novel attributes ("DateFest" from Eastwick, Smith, et al., 2019). The challenge was modifying this paradigm so that it could believably be applied to attributes that might characterize someone's own romantic partner.

DateFest

DateFest is based on BeanFest, a paradigm that prompts people to form attitudes towards novel attitude objects (i.e., beans; Fazio et al., 2004, 2015). Broadly speaking, DateFest's experimental conditions prompt people to prefer a novel attribute to varying degrees (Eastwick, Smith, et al., 2019) through a game where they need to decide whether or not to go on a date with hypothetical partners. Depending on the condition, the game links the presence of the novel attribute either strongly or weakly with the rewards and costs associated with going on dates.

The structure of the game contains a revealed (or "functional") preference that participants experience, which subsequently affects their own stated (or "summarized") preference. In essence, a stronger attribute-reward association in the game causes participants to say they "like" the attribute more (Ledgerwood et al., 2018).

An initial iteration of DateFest involved fantastical attributes (called "Melb" and "Flobe") characterizing alien inhabitants of another planet. However, to investigate downstream consequences of ideal partner preferences for people's actual romantic relationships in the current project, DateFest needed to be modified to manipulate ideals for traits that participants' partners could conceivably possess. Therefore, we created a version of DateFest in which (a) participants evaluated other humans as dating partners (i.e., faces from the Chicago Face Database; Ma et al., 2015), and (b) the manipulated attribute was a real feature of the faces of the potential dating partners.

Goals of the Current Research

Our first goal was to use best practices (e.g., a high-powered design, preregistration) to convincingly demonstrate that an experimental procedure could cause shifts (of a meaningful effect size) in ideal partner preferences for a real (i.e., not alien) attribute (Studies 1-3). Our second goal was to test whether this manipulation has downstream consequences consistent with the various theoretical perspectives described above. Specifically, we examined 6 research questions about the causal implications of ideal partner preferences[13]:

RQ1. Motivated projection, primary DV (Studies 1-3): Do ideal partner preferences cause people to perceive a current romantic partner to be higher on the relevant attribute?

RQ2. Motivated projection, secondary DV (Studies 1-3): Do ideal partner preferences cause people to report higher relationship satisfaction?

RQ3a. Ideal-partner preference matching (Studies 1-3): Do ideal partner preferences cause people to report higher relationship satisfaction to the extent that they perceive their partner to possess the relevant attribute? (i.e., Condition × Perception of attribute = Satisfaction).

RQ3b. Moderated motivated projection (Studies 1-3): Do ideal partner preferences cause people to perceive a current romantic partner to be higher on the relevant attribute to the

[13] In the preregistrations, RQ5 was listed as RQ6, RQ4 was RQ5, and RQ3a and 3b were RQ3 and RQ4. We made these changes to this manuscript to address a reviewer's concern that the ideal-partner preference matching (3a) and the moderated motivated projection (3b) analyses are two ways of testing the same moderational effect.

extent that they are currently satisfied? (i.e., Condition × Satisfaction = Perception of attribute; see Footnote 1).

RQ4. Situation selection (Studies 1-3): Do ideal partner preferences cause people to desire to join a website that features partners who are high on the relevant attribute?

RQ5. Perceiver effects (Studies 2-3): Do ideal partner preferences cause people to boost their estimates of the extent to which the attribute characterizes friends, strangers, and disliked others?

Study 1

Study 1 aimed to establish a paradigm for manipulating ideal partner preferences. For this purpose, there were two between-subjects conditions: strong vs. weak functional preference for the focal trait "Reditry" (which was in reality the extent to which dating partner seemed youthful or babyfaced). Specifically, in the strong functional preference condition, Reditry was strongly associated with the likelihood of going on enjoyable dates, and in the weak functional preference condition, the association was only modest. That is, participants' experiences with higher levels of Reditry was more positive in the strong than the weak condition.

The study used a modified DateFest paradigm and an unfamiliar name for youthfulness (Reditry) to recreate the experience of forming a preference in the first place and to circumvent participants' existing beliefs about youthfulness (da Silva Frost et al., 2022). As a secondary goal, we wanted to start investigating the effects consistent with the theoretical approaches described above. Thus, after the manipulation, we tested several of the causal implications of ideal partner preferences (i.e., Research Questions 1-4).

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The data, analysis code, materials, and internal (i.e., on a lab server) preregistration can be found at https://osf.io/mptqk/?view_only=06b8f1fcbd0843c5b24e2b3353353a36. (In Studies 2 and 3, the preregistration was externally posted to the relevant OSF folder before running the study.)

Participants and Power

Our final sample consisted of *N* = 403 (365 females, 37 males, and 1 person who chose another option; Mage = 27.7, SD = 4.5; 61.5% White, 13.6% Black, 7.8% Asian or Pacific Islander, 7.7% Hispanic or Latino, 5.2% Biracial or Multiracial, 2.3% Native American and 1.7% other). Of these, *n* = 191 were randomly assigned to the strong functional preference condition and *n* = 212 were assigned to the weak functional preference condition. All these participants were between 18 and 35 years old, were primarily attracted to men, and reported being involved in a romantic relationship (*M* length = 58.7 months); participants had to be involved in a relationship to be eligible for the current study so that the primary DVs of interest (e.g., estimating the Reditry of one's partner, relationship satisfaction) would make sense.

Our final sample of 403 (after exclusions) came from a larger sample of 654 participants who completed the study online through the Mturk platform. We set and recorded *a priori* to collect at least *N* = 580 to arrive at a final sample of 400; we collected many more than our target *N* = 400 because we anticipated (based on prior experience) a high number of attention-check exclusions (~30%). In total, *n* = 251 participants were excluded (*n* = 112 failed the "other" attention check, *n* = 181 failed the Winograd-like schema check, and *n* = 42 failed both). We targeted a sample size of *N* = 400, which provides the ability to detect an effect as small as *d* = 0.28 with 80% power.

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 men were able to proceed, according to the inclusion criteria. Participants then completed the Date-Fest task.

Afterwards, participants completed the ideal preference for Reditry manipulation check items and an estimate of the partner's Reditry; the order of these two measures was counterbalanced. They then completed (in order) the relationship satisfaction measure, the sense of humor measures (ideals, and ratings of the partner), and the situation selection measures.

Covariation detection task. Participants played a game called DateFest (Eastwick, Smith, et al., 2019) in which the goal was to gain points by making rewarding dating decisions. Participants were told that they were going to a party where they would meet 24 party guests and that they must decide whether to go on a date with each one. Participants learned that some of the dates would be good experiences whereas others would be bad experiences, and they needed to figure out which guests would lead to good vs. bad experiences in order to gain points. Choosing to go on the dates that were a bad experience (which was true for 12 out of 24 guests) caused participants to lose 10 points, and choosing dates that were a good experience (the other 12 guests) caused them to gain 10 points.

Participants only lost or gained points when they chose to go on dates. If they chose not to go on a date with a guest, they neither gained nor lost any points. To ensure participants had access to the same amount of information regardless of how many dates they chose to go on, they learned whether they would have gained or lost points if they had gone on the date.

Therefore, the task was designed to induce participants to form more positive evaluations when the traits were associated with gaining (vs. losing) points.

Manipulating functional preference strength. We manipulated the functional (i.e., induced) preference for Reditry by varying the extent to which Reditry was associated with gaining vs. losing points in the covariation detection task. In the weak functional preference condition, the party guests that caused participants to earn versus lose points had very similar average values of Reditry: The average Reditry of the good dates was 81 and the average Reditry of the bad dates was 58 (Figure 9). In other words, Reditry was only a modest predictor of the extent to which the date would be a good experience and generate points. In the strong functional preference condition, the guests that caused participants to gain versus lose points had very different average values of Reditry: The average Reditry of the good dates was 90 and the average Reditry of the bad dates was 48 (Figure 10). In other words, Reditry was a strong predictor that the date would be a good experience and generate points. The overall average Reditry of all 24 potential mates was held constant across the weak functional preference and strong functional preference conditions (i.e., the average Reditry was always 69). Finally, both conditions used the same faces; the condition manipulation simply shifted which guests led to good or bad dates.

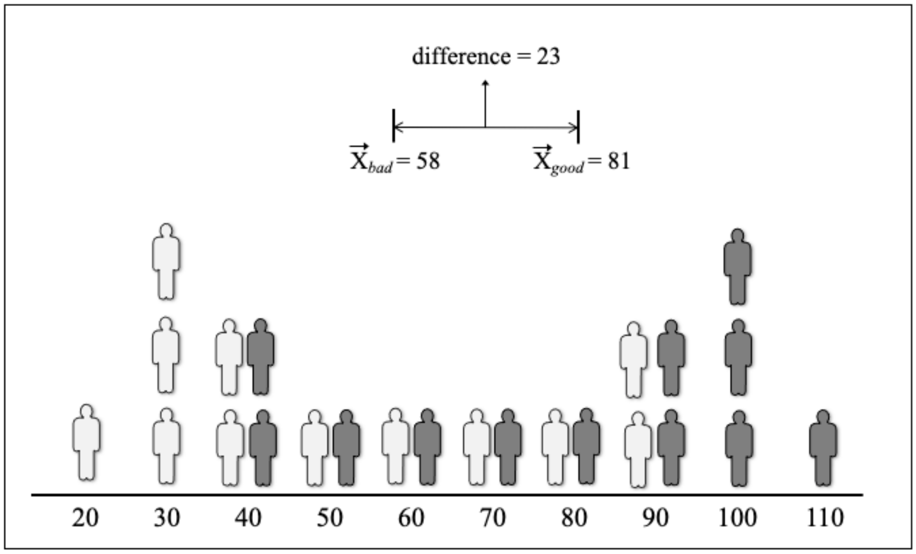


Figure 9. Weak functional preference for Reditry condition in Study 1. Each person represents a party guest and numbers in the bottom indicate their (approximate) level of Reditry. The people painted in black represent party guests with whom participants had good dates (and earned them 10 points), and the people painted in white represent party guests with whom participants had bad dates (and cost them 10 points).

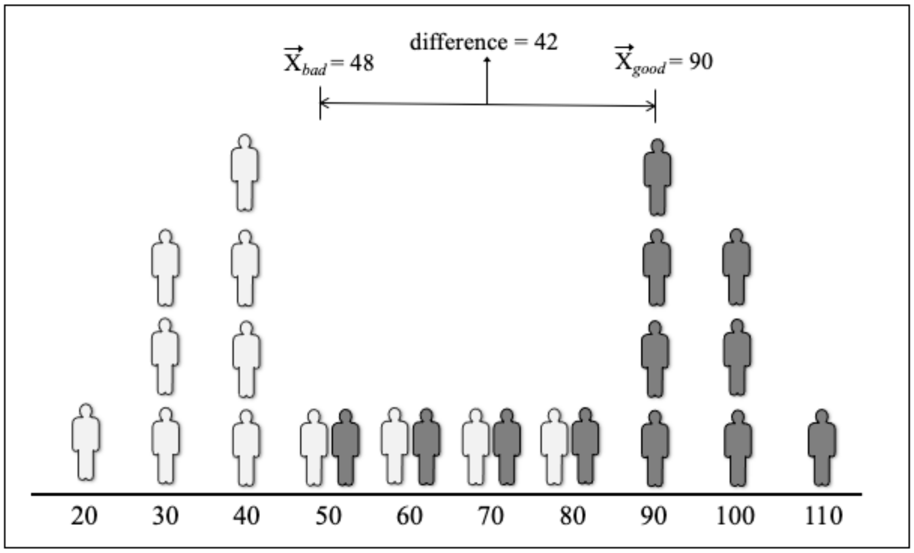
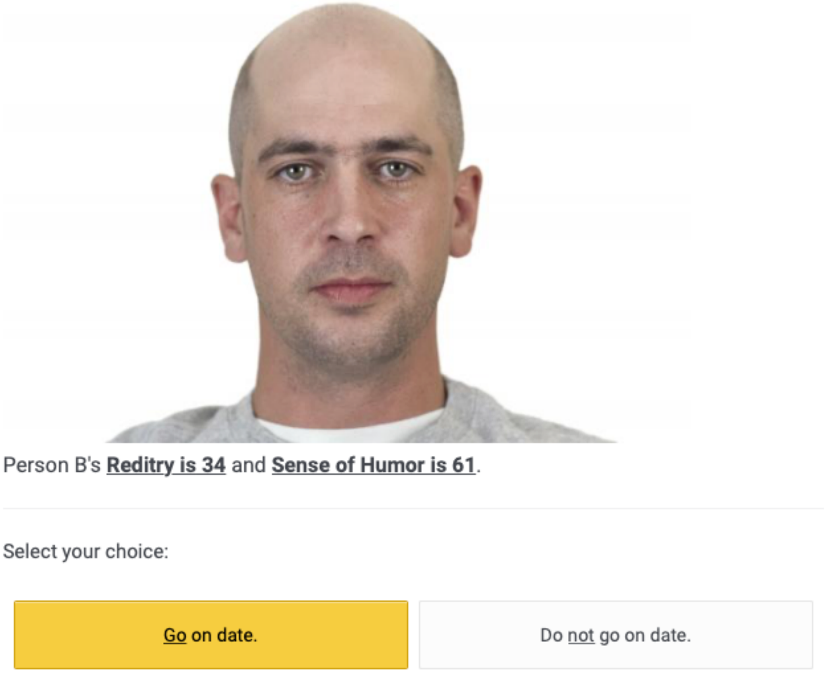


Figure 10. Strong functional preference for Reditry condition in Study 1. Each person represents a party guest and numbers in the bottom indicate their (approximate) level of Reditry. The people painted in black represent party guests with whom participants had good dates (and earned them 10 points), and the people painted in white represent party guests with whom participants had bad dates (and cost them 10 points).



Person B's **Reditry is 34** and **Sense of Humor is 61**.

Select your choice:

Figure 11. Example of a trial. Notice that for each party guest, participants learned the amount of Reditry and sense of humor.

Besides Reditry, participants had to track an additional control trait (sense of humor); these values were unrelated to the manipulation. Participants saw each party guest's face and learned how much Reditry and sense of humor the guest had before making a "date" vs. "do not date" decision (Figure 11).

Reditry was babyfacedness as measured in the CFD norming data (Ma et al., 2015); we selected 24 faces from this database to use as stimuli (babyfacedness *M* = 2.55, *SD* = 0.70, range = 1 to 7) and we algebraically transformed these values to range from 1 to 120. We did not manipulate the functional preference strength of sense of humor, and these values (which also ranged from 1 to 120) were not connected to features of the face. To ensure that sense of humor was equally likable across both the weak and strong Reditry functional preference conditions, the good dates always had an average sense of humor of 85 and the bad dates always had an average sense of humor of 53. Reditry and sense of humor levels were chosen so that the two traits did not correlate with each another (*r*s ranged from = -.01 to .02) within the set of liked targets and within the set of disliked targets; see Supplemental Materials for details on stimuli.

Ideals for Reditry (manipulation check). After playing DateFest, participants responded to the following four questions, which comprised the ideal partner preference for Reditry manipulation check: “How important is Reditry to you in a romantic partner?”, “How much do you value Reditry in a romantic partner?,” “How desirable is Reditry to you in a romantic partner?,” and “To what extent does Reditry characterize your ideal romantic partner?” on scales from 1 (not at all) to 9 (extremely). These four items were highly reliable (α = .97) and were thus averaged to form a scale reflecting participants’ ideal preference for Reditry. To test for discriminant validity, participants also responded to these same questions about sense of humor (α = .94).

Participants’ estimate of partner’s attributes. In order to assess participants’ estimate of their partners’ attributes, we added the following items: “If you had to guess, how much Reditry do you think your current romantic partner has?” (used in RQs 1, 3a, and 3b) and “If you had to guess, how much sense of humor do you think your current romantic partner has?” (used for discriminant validity) on scales from 1 (very little Reditry/sense of humor) to 12 (A lot of Reditry/sense of humor).

Relationship satisfaction. We averaged the five satisfaction items from the Investment Model Scale (Rusbult et al., 1998): “I feel satisfied with our relationship”, “My relationship is much better than others’ relationships”, “My relationship is close to ideal”, “Our relationship makes me very happy” and “Our relationship does a good job of fulfilling my needs for intimacy, companionship, etc." (used in RQs 2, 3a, and 3b). Items were measured on a scale from 1 (do not agree at all) to 9 (completely agree); α = .96.

Situation selection items. We included items assessing situation selection in all three studies (RQ4). To assess participants' interest in entering a situation with potential partners high in Reditry, participants read the following prompt: "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. How interested are you in the website that would only include partners high in Reditry?" They responded on a 9-point Likert-type scale (from 1 = *not at all interested* to 9 = *very interested*). Participants then completed an identical item about sense of humor for discriminant validity.

Attention checks. We included two attention checks. The first one was a question in the demographics section that instructed participants to select "other" instead of their actual region of origin. Participants who did not follow the instructions were excluded from the analysis. The second attention check 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, Davis, & Morgenstern, 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., “unfortunately”), as coded by a researcher blind to the study results.

Results

Preference for Reditry (Manipulation Check)

A planned independent samples t-test examined whether participants reported greater liking for Reditry in the strong than the weak functional preferences condition. Participants indicated a higher ideal partner preference for Reditry in the strong (*M* = 6.44, *SD* = 1.84) than the weak (*M* =5.71, *SD* = 2.03) condition, *t*(401) = 3.80, *p* < .001, *d* = 0.38, 95% CI [0.18, 0.58]. In other words, the manipulation caused participants to boost their ratings of the importance of Reditry in an ideal romantic partner.

Research Questions

RQ1: Estimate of Reditry in a current partner. Consistent with the motivated perception account, participants believed that their current partners had more Reditry in the strong (*M* = 8.73, *SD* = 2.19) than the weak (*M* = 8.05, *SD* = 2.35) functional preferences condition, *t*(401) = 2.98, *p* = .003, *d* = 0.30, 95% CI [0.10, 0.49]. In other words, when the experimental manipulation caused participants to feel more positively about Reditry in an ideal romantic partner, they inferred that their partners had higher Reditry.

RQ2: Relationship satisfaction with a current partner. Participants’ reports of their relationship satisfaction were slightly higher in the strong (*M* = 7.70, *SD* = 1.52) than the weak (*M* = 7.41, *SD* = 1.66) functional preferences condition, but this difference was only marginally significant, *t*(401) = 1.81, *p* = .072, *d* = 0.18, 95% CI [-0.02, 0.38]. We cannot conclude that ideal partner preferences causally boost downstream judgments of relationship satisfaction.

RQ3a: Ideal partner preference-matching (trait weighting). To test the preference-matching (trait weighting) account, we conducted a regression using condition (coded -1 = weak, 1 = strong), partner Reditry (standardized), and their interaction to predict relationship satisfaction (also standardized). Results are depicted in Figure 12. The main effect of condition was not significant, $\beta = 0.03$, $t(399) = 0.60$, $p = .552$, and the main effect of partner Reditry was significant, $\beta = 0.42$, $t(399) = 9.16$, $p < .001$. Critically, although the interaction term was in the predicted direction (i.e., positive), it was not significant, $\beta = 0.04$, $t(399) = 0.95$, $p = .341$. In other words, there was no causal evidence that participants placed a stronger weight on Reditry in their evaluations of their partners.

RQ3b: Moderated motivated perception. To test the moderated motivated perception account, we conducted a regression using condition (coded -1 = weak, 1 = strong), satisfaction (standardized), and their interaction to predict partner Reditry (also standardized). Results are depicted in Figure 13. The main effect of condition was significant as in the basic t-test in RQ1, $\beta = 0.11$, $t(399) = 2.42$, $p = .016$, and the main effect of satisfaction was significant, $\beta = 0.42$, $t(399) = 9.18$, $p < .001$. As with RQ3a, the interaction was not significant, $\beta = 0.05$, $t(399)= 1.10$, $p = .270$.

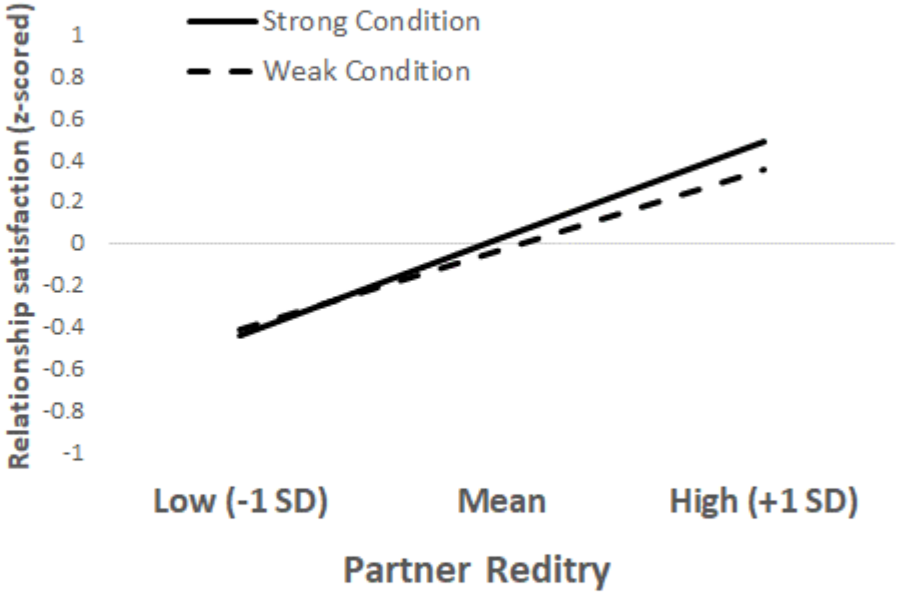


Figure 12. Regression results for RQ3a. The slope of Reditry predicting satisfaction did not differ by condition, which does not support the causal account for preference-matching.

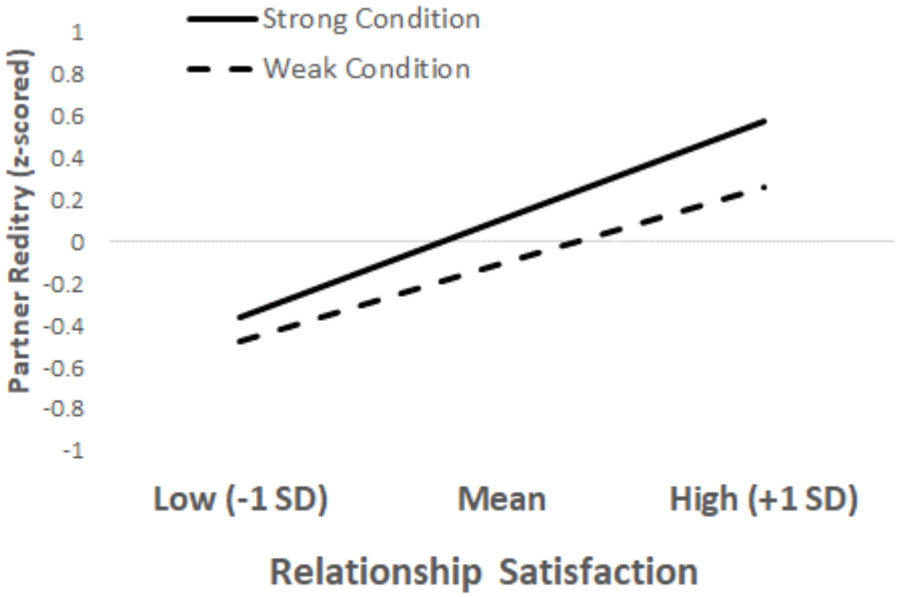


Figure 13. Regression results for RQ3b. The slope of satisfaction predicting Reditry did not differ by condition, which does not support the moderated motivated perception account.

RQ4: Situation selection. Participants reported greater interest in joining a website with potential partners high in Reditry in the strong (*M* = 6.13, *SD* = 2.16) than the weak (*M* = 5.46, *SD* = 2.48) functional preferences condition, *t*(401) = 2.87, *p* = .004, *d* = 0.29, 95% CI [0.09, 0.48]. In other words, the experimental manipulation boosted participants' interest in selecting into a situation containing partners high in Reditry. Figure 14 presents the effect sizes associated with the manipulation check and the RQs.

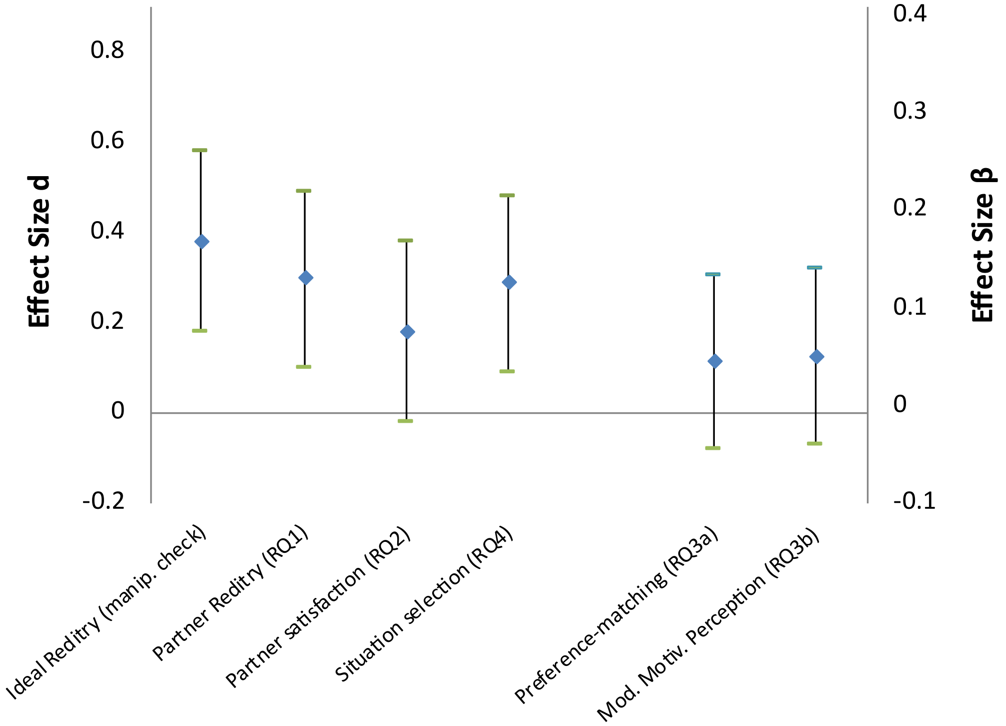


Figure 14. Effect sizes for Study 2. Left (4) effect sizes are *d*s; right (2) effect sizes are *β*s. Axes are scaled to be equivalent, $\beta = d \div \sqrt{(d^2 + 4)}$. Bars are 95% CIs.

Sense of Humor (Control) Questions

We also tested whether condition affected sense of humor, a trait that was unrelated to the manipulation. Participants' ideals for sense of humor did not differ across the strong (*M* = 8.03, *SD* = 1.03) vs. weak (*M* =8.08, *SD* = 1.01) functional preference conditions, *t*(401) = -0.46, *p* = .645, *d* = -0.05, 95% CI [-0.24, 0.15].

Not surprisingly, participants' estimates of the extent to which their partner has a sense of humor did not differ across the strong (*M* = 9.85, *SD* = 1.76) vs. weak (*M* = 9.59, *SD* = 2.05) functional preference conditions, *t*(401) = 1.35, *p* = .177, *d* = 0.14, 95% CI [-0.06, 0.33]. Similarly, participants' interest in joining a dating website with potential partners high in sense of humor did not differ across the strong (*M* = 7.28, *SD* = 1.83) vs. weak (*M* = 7.21, *SD* = 2.00) functional preferences conditions, *t*(401) = 0.37, *p* = .715, *d* = 0.04, 95% CI [-0.16, 0.23].

Collectively, these results suggest that our manipulation of Reditry affected only the DVs related specifically to Reditry, not the control attribute sense of humor.

Mediation of partner Reditry on RQ2

We ran an exploratory analysis suggested by a reviewer that examined whether the effect of condition on satisfaction was mediated by perceptions of the partner's Reditry. Using the component approach recommended by Yzerbyt et al. (2018), we found evidence for mediation: Results showed that path a (i.e., condition on partner Reditry; β = 0.15, p =.003) and path b (i.e., partner Reditry on satisfaction controlling for condition; β = 0.42, p <.001) were both significant. The direct effect (i.e., condition on satisfaction; β = 0.03, p =.533) was not significant, and the indirect effect (β = 0.06, 95% CI [.02, .11]) accounted for most of the total effect. In other words, despite the fact that the manipulation did not affect satisfaction directly, our results are consistent with the Murray et al. (1996a, 1996b) account that ideals cause people to elevate their perception of the partner's traits, which is in turn associated with higher satisfaction.

Discussion

In Study 1, we used the modified DateFest paradigm to manipulate ideal partner preferences for a novel attribute. In the strong functional preference condition, Reditry—which was actually youthfulness—was strongly associated with the rewards that participants experienced on their dates. In the weak functional preference condition, Reditry was only modestly connected to the rewards that participants experienced.

Post-game, participants in the strong (vs. weak) functional preference condition reported a stronger preference for Reditry in a romantic partner, with a moderate effect size—a demonstration that the manipulation worked as intended. Critically, participants in the strong (vs. weak) functional preference condition estimated that their partners had higher amounts of Reditry, in line with the motivated projection account (RQ1). Participants in the strong (vs. weak) functional preference condition were more interested in joining a website with potential partners high in Reditry (RQ4), with a similarly strong effect size. None of our control (i.e., sense of humor) variables were significant.

The difference in relationship satisfaction between conditions (RQ2) was marginally significant. This suggestive but inconclusive result led us to power the next study so that we would be likely to detect this difference, if it exists. Thus, in Study 2, we increased both sample size and manipulation strength.

Neither the ideal-partner-preference matching (i.e., condition × partner Reditry = satisfaction interaction for RQ3a) nor the moderated motivated projection accounts (condition × satisfaction = partner Reditry interaction for RQ3b) were significant. These results are not consistent with a causal account where ideals cause higher relationship satisfaction if partners possess the relevant attribute (i.e., the trait weighting model of preference-matching), nor is it consistent with causal effect of ideals on participants' perceptions of their partner being stronger if they are especially satisfied. Nevertheless, both effect sizes trended in the expected direction, so we tested these hypotheses again in Study 2 with a stronger manipulation and larger sample.

Study 2

In Study 2, our main objective was to replicate Study 1 using both (a) a stronger manipulation and (b) a larger sample size (da Silva Frost & Ledgerwood, 2020; Ledgerwood & Shrout, 2011). As in Study 1, Study 2 investigated RQ1-4. We also tested the perceiver effect account (RQ5) by examining whether the effect of condition on the estimate of the partner's trait was specific to romantic partners or a more general pattern that extended across different kinds of targets. For this purpose, we added items to assess participants' estimates of Reditry in themselves, a friend, a stranger, and a disliked other.

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The data, analysis code, materials, and external (i.e., posted on OSF before data collection) preregistration can be found at https://osf.io/mptqk/?view_only=06b8f1fcbd0843c5b24e2b3353353a36. In this study, the manipulation check, RQ1, RQ2, RQ3a, RQ3b and RQ5 (with the Bonferroni-Holm correction; Holm, 1979) were all preregistered. Three other preregistered analyses are included in the Supplemental Materials.

Participants and Power

In Study 1, the effect of manipulation on relationship satisfaction with $N = 403$ was $d = .18$ [-.02, .38]. If this is a reasonable effect size estimate for this effect, we would be powering our study at 90% with a sample size of $N = 1300$. Because we were expecting a high attention-check failure rate, we preregistered that we would collect *N* = 1857 (expecting to have at least 1300 after the estimated exclusion rate of ~30%). We also preregistered that, if after exclusions and before running any analyses we have a sample size of less than 1300, we would compute the exclusion rate ("e") and collect n = (1300-current n)/(1-e) additional participants.

Following this procedure, we ended up with *N* = 1855 who met the inclusion criteria used in Study 1 (i.e., between 18 and 35 years old, were primarily attracted to men, and in a romantic relationship). We recorded *a priori* three exclusion criteria. The first two were identical to the previous studies: we had an attention check in which participants were requested to select the option "other", and a Winograd-like schema check in which participants answered open ended questions about a short story and were excluded for nonsensical answers. Besides these, we also requested that participants to write the initials of their romantic partners for later reference in the study, and we excluded participants who provided nonsensical answers (e.g., "2", "nothing") Given that the manipulation used male faces, we also decided to exclude participants who used female names for their romantic partner (e.g. "Laura," "Susan"). In total, *n* = 216 participants were excluded (*n* = 199 failed the "other" attention check, *n* = 33 failed the Winograd-like schema check and *n* = 11 failed the initials check). Thus, after exclusions our final sample was *N* = 1639 (1536 women, 97 men, and 6 people who chose another option; M_{age} = 28.5, *SD* = 4.13; 63.9% White, 10.5% Black, 6.7% Asian or Pacific Islander, 10.4% Hispanic or Latino, 5.4% Biracial or Multiracial, 0.9% Native American and 2.2% other). Of these, *n* = 836 were randomly assigned to the strong functional preference condition and *n* = 803 were assigned to the weak functional preference condition, and they had been in a relationship of 64.8 months on average.

Procedure and Materials

The procedure and measures in this study were identical to Study 1, except for two changes. First, to create a stronger manipulation, we made the difference between the two functional preference conditions more extreme (see Figures 15 and 16). That is, in Study 1, the difference between the good and bad dates was 23 Reditry points in the weak functional preferences condition and 42 points in the strong condition. In this study, the difference between the good and bad dates was 7 Reditry points in the weak functional preferences condition and 53 points in the strong condition. As in Study 1, the amount of total Reditry was the same across the two conditions (69.5 on average), the association between sense of humor and the good vs. bad date difference did not differ by condition, and the two traits did not correlate with each another (*r*s ranged from = -.06 to .04) within the set of liked targets and within the set of disliked targets. The preference for Reditry (α = .98), preference for sense of humor (α = .94), and relationship satisfaction (α = .96) measures were the same as in Study 1.

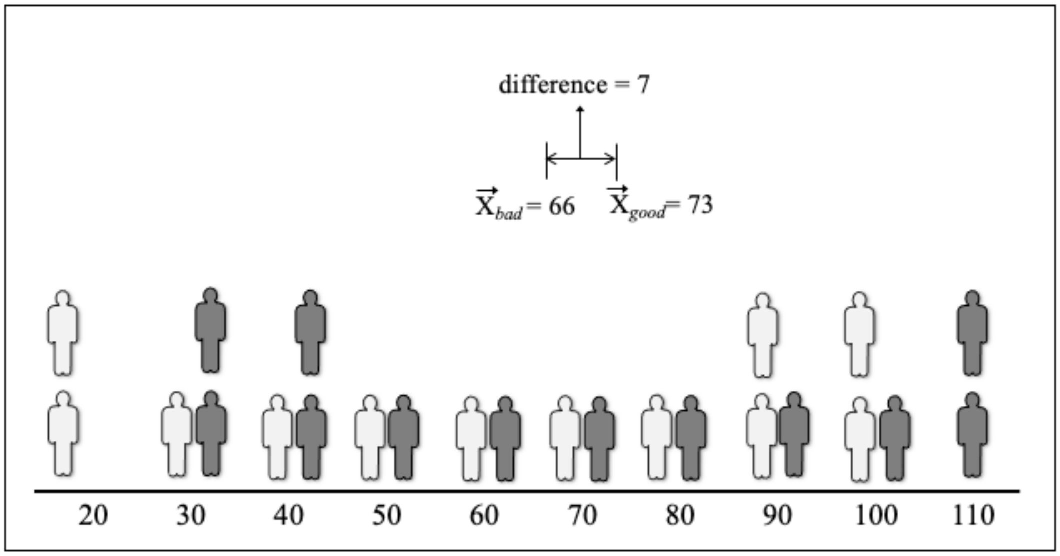


Figure 15. Weak functional preference for Reditry condition in Study 2. Each person represents a party guest and numbers in the bottom indicate their approximate level of Reditry, in blocks of 10. The people painted in black represent party guests with whom participants had good dates (and earned them 10 points), and the people painted in white represent party guests with whom participants had bad dates (and cost them 10 points).

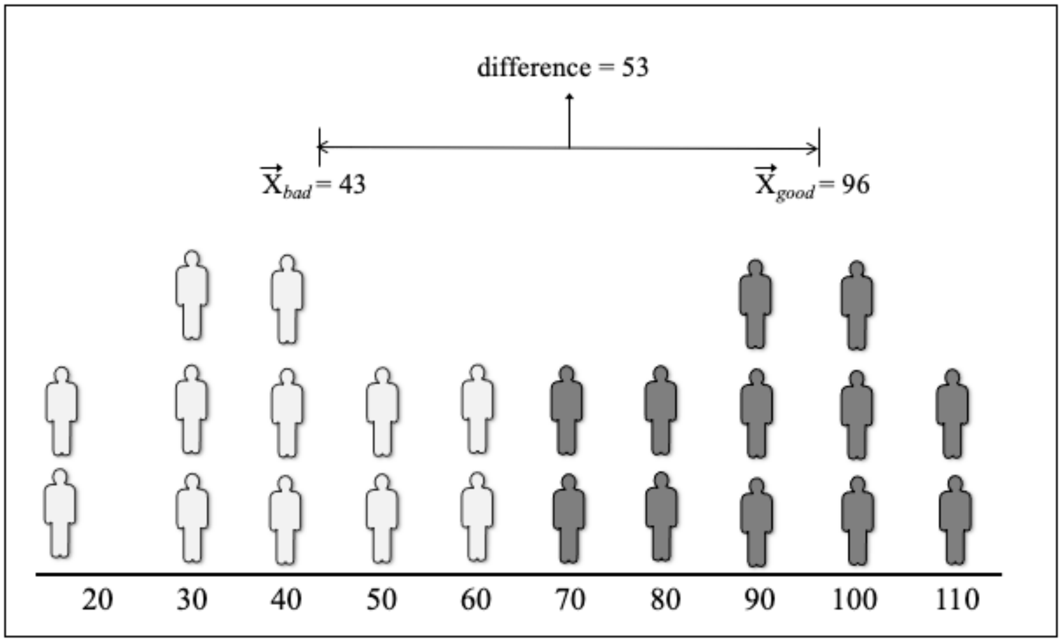


Figure 16. Strong functional preference for Reditry condition in Study 2. Each person represents a party guest and numbers in the bottom indicate their approximate level of Reditry, in blocks of 10. The people painted in black represent party guests with whom participants had good dates (and earned them 10 points), and the people painted in white represent party guests with whom participants had bad dates (and cost them 10 points).

Second, to test RQ5, we added other targets besides romantic partner. When entering the study, participants were prompted to provide initials of their romantic partner, a friend, and a disliked person. Then, after completing the main measures about their romantic partner, we also asked them to estimate the amount of Reditry in themselves, the friend and the disliked person. To do this, we simply replaced "romantic partner" with the appropriate target, accompanied by the initials that participants chose at the beginning of the study. We asked participants to estimate the amount of Reditry in a stranger as well, by prompting them to think of the most recent male stranger they had an encounter with. The order of these four additional targets was presented randomly.

Results

Preference for Reditry (Manipulation Check)

As in Study 1, participants indicated a stronger preference for Reditry in the strong (*M* = 6.55, *SD* = 1.93) than the weak (*M* =5.06, *SD* = 1.96) functional preferences condition, *t*(1636) = 15.56, *p* < .001, *d* = 0.77, 95% CI [0.67, 0.87]). In other words, the manipulation caused participants to boost their ratings of the importance of Reditry in an ideal romantic partner. The magnitude of this manipulation was approximately double that of Study 1, which is consistent with our use of a stronger manipulation.

Research Questions

RQ1: Estimate of Reditry in a current partner. Consistent with the motivated perception account, participants believed that their current partners had more Reditry in the strong (*M* = 8.84, *SD* = 2.13) than the weak (*M* = 7.24, *SD* = 2.44) functional preferences condition, *t*(1637) = 14.15, *p* < .001, *d* = 0.70, 95% CI [0.60, 0.80]). Once again, when the experimental manipulation caused participants to feel more positively about Reditry in an ideal romantic partner, they inferred that their partners had higher Reditry; the effect size in this study was very large.

RQ2: Relationship satisfaction with a current partner. Participants' reports of their relationship satisfaction with their current partner did not differ between the strong (*M* = 7.55, *SD* = 1.63) and the weak (*M* = 7.51, *SD* = 1.64) functional preferences conditions, *t*(1637) = 0.50, *p* = .616, *d* = 0.03, 95% CI [-0.07, 0.12]). That is, even with a very powerful manipulation and a large sample size, this study was unable to provide any evidence that the manipulation caused participants to boost their relationship satisfaction.

RQ3a: Ideal partner preference-matching (trait weighting). To test the preference-matching (trait weighting) account, as in Study 1, we conducted a regression using condition (coded -1 = weak, 1 = strong), partner Reditry (standardized), and their interaction to predict relationship satisfaction (also standardized). Results are depicted in Figure 17. The main effect of condition was significant (and actually negative), $\beta = -0.11$, $t(1635) = -4.54$, $p < .001$, and the main effect of Reditry was significant, $\beta = 0.38$, $t(1635) = 15.35$, $p < .001$. Importantly, the condition × Reditry interaction was significant, $\beta = 0.16$, $t(1635) = 6.80$, $p < .001$. That is, the matching effect emerged: When participants were experimentally induced to feel more (vs. less) positively about Reditry, the extent to which they felt their partner had Reditry was a stronger predictor of their relationship satisfaction (i.e., the solid black line is steeper than the dashed line in Figure 17).

RQ3b: Moderated motivated perception. To test the moderated motivated perception account, as in Study 1, we conducted a regression using condition (coded -1 = weak, 1 = strong), satisfaction (standardized), and their interaction to predict partner Reditry (also standardized). Results are depicted in Figure 18. The main effect of condition was significant as in the basic t-test in RQ1, $\beta = 0.33$, $t(1635) = 14.94$, $p < .001$, and the main effect of satisfaction was significant, $\beta = 0.32$, $t(1635) = 14.49$, $p < .001$. Importantly, in parallel with RQ3a, we also found a significant condition × satisfaction interaction, $\beta = 0.11$, $t(1635) = 4.81$, $p < .001$. That is, the experimental manipulation was more effective at boosting participants' perceptions of their partner's Reditry to the extent that they were satisfied with the relationship.

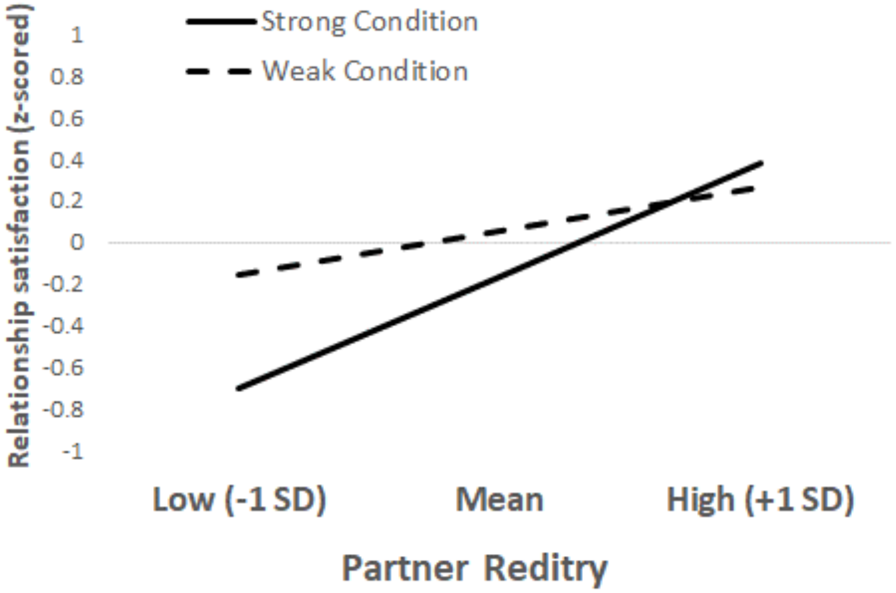


Figure 17. Regression results for RQ3a. The slope of Reditry predicting satisfaction did differ by condition, which supports the causal account for preference-matching.

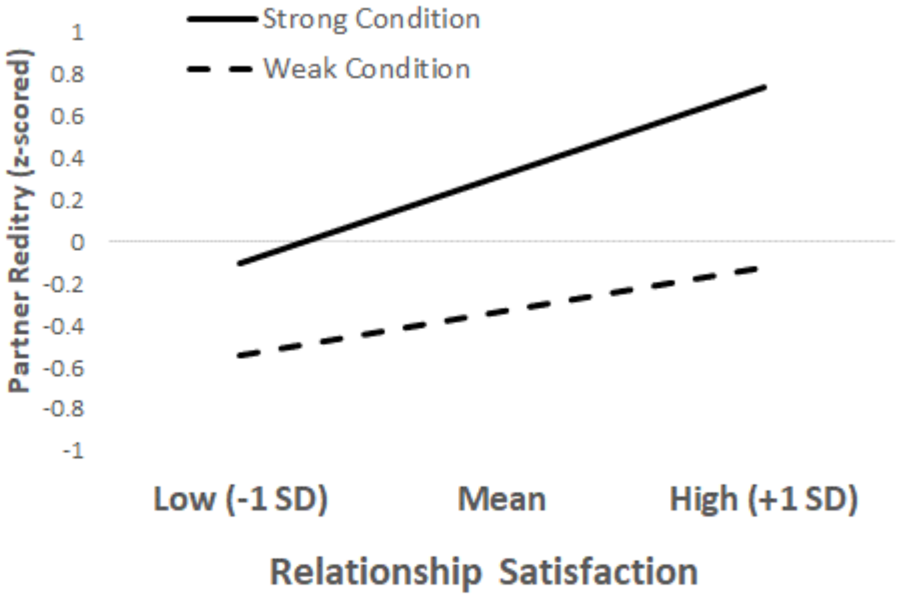


Figure 18. Regression results for RQ3b. The slope of satisfaction predicting Reditry did differ by condition, which supports the moderated motivated perception account.

RQ4: Situation selection. Participants reported greater interest in joining a website with potential partners high in Reditry in the strong (*M* = 5.97, *SD* = 2.38) than the weak (*M* = 4.62, *SD* = 2.52) functional preferences condition, *t*(1636) = 11.17, *p* < .001, *d* = 0.55, 95% CI [0.45, 0.65]. That is, once again, the manipulation boosted participants' interest in selecting into a situation containing partners high in Reditry.

Figure 19 presents the effect sizes associated with the manipulation check and the RQs; relative to Study 1, the effect sizes are larger (due to the stronger manipulation) and the SEs are smaller (due to the larger *N*).

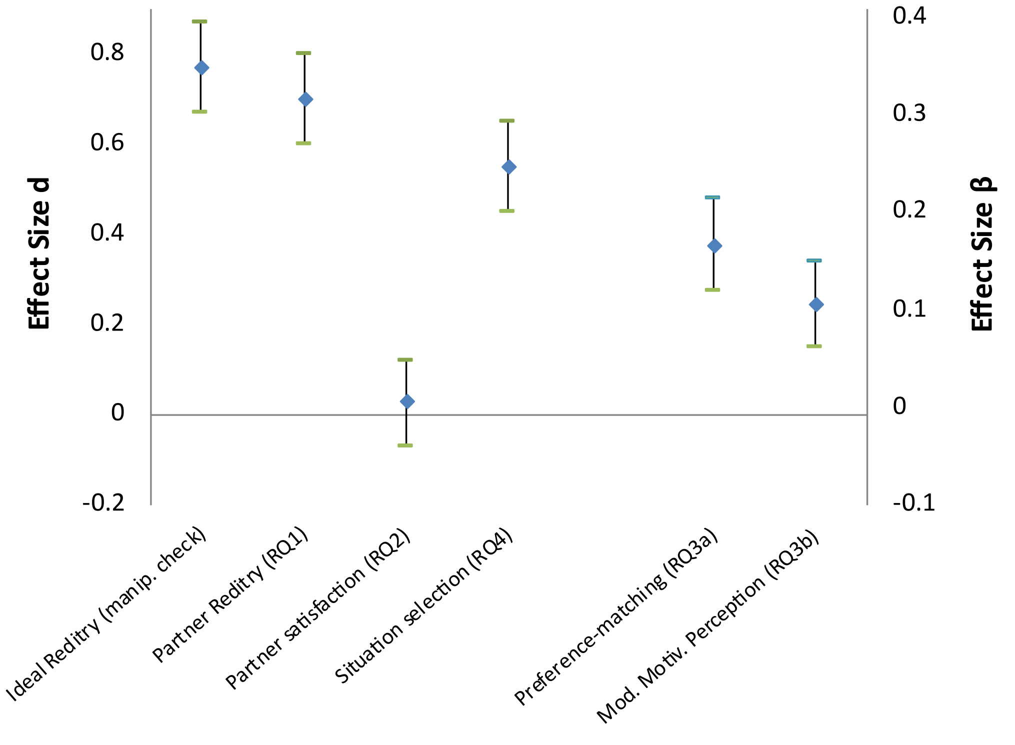


Figure 19. Effect sizes for Study 2. Left (4) effect sizes are *d*s; right (2) effect sizes are *β*s. Axes are scaled to be equivalent, $\beta = d \div \sqrt{(d^2 + 4)}$. Bars are 95% CIs.

RQ5: Perceiver effects. Did the manipulation also cause participants to boost the extent to which they thought *other* people possessed Reditry, besides their romantic partner? Figure 20 illustrates the effect sizes of the strong vs. weak condition on partner Reditry (RQ1), as well as ratings of Reditry in the self, a friend, a stranger, and a disliked other. Ratings of Reditry in the self were higher in the strong (*M* = 8.01, *SD* = 2.04) than the weak (*M* = 6.75, *SD* = 2.29) functional preferences condition, *t*(1637) = 11.78, *p* < .001, *d* = 0.58, 95% CI [0.48, 0.68]; ratings of Reditry in a friend were higher in the strong (*M* = 7.89, *SD* = 2.20) than the weak (*M* = 6.75, *SD* = 2.37) functional preferences condition, *t*(1622) = 10.07, *p* < .001, *d* = 0.50, 95% CI [0.40, 0.60]; ratings of Reditry in a stranger were higher in the strong (*M* = 6.28, *SD* = 2.07) than the weak (*M* = 5.98, *SD* = 2.10) functional preferences condition, *t*(1637) = 2.87, *p* = .004, *d* = 0.14, 95% CI [0.05, 0.24]; and ratings of Reditry in a disliked other were actually *lower* in the strong (*M* = 3.77, *SD* = 2.54) than the weak (*M* = 4.92, *SD* = 3.26) functional preferences condition, *t*(1631) = 7.95, *p* < .001, *d* = -0.39, 95% CI [-0.49, -0.30]. All these differences remained significant after a Bonferroni-Holm (Holm, 1979) correction, as preregistered.

In short, the manipulation affected participants' ratings of the Reditry of all targets, as predicted by the perceiver effects account. Furthermore, it is notable that the effect was strongest on ratings of the partner and weakest on ratings of a stranger; effects for the self, friend, and disliked other were intermediate in size.

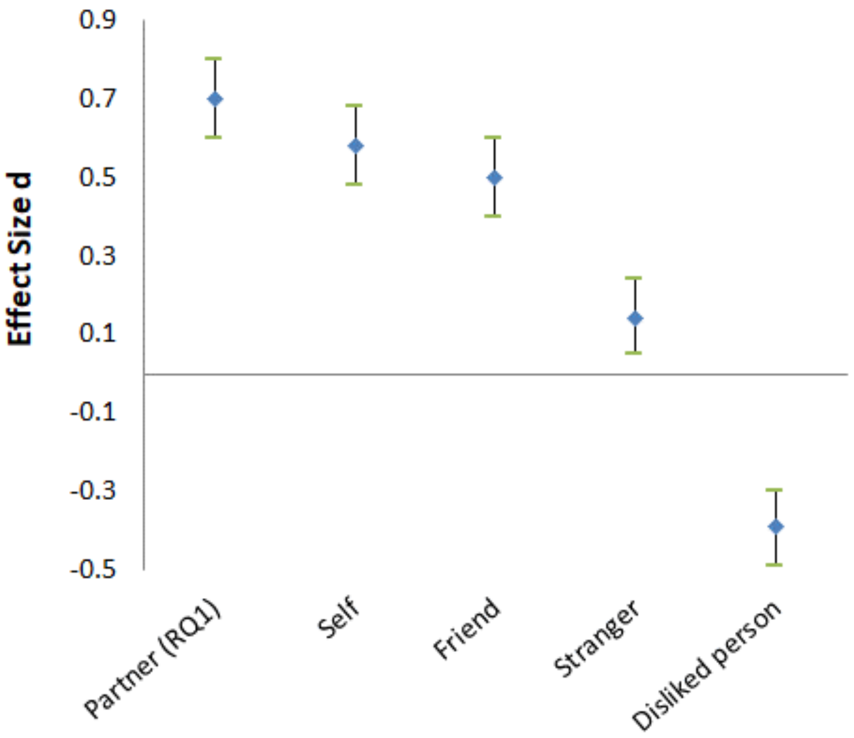


Figure 20. Effect sizes for RQ5, Study 3. Effect sizes (*d*) illustrate the effect of the manipulation on participants' ratings of the Reditry of five different targets. Bars are 95% CIs.

Sense of Humor (Control) Questions

Again, we tested whether condition affected sense of humor. Participants' ideals for sense of humor were actually lower in the strong (*M* =7.81, *SD* = 1.25) than the weak (*M* = 7.99, *SD* = 1.08) functional preferences condition, *t*(1637) = 3.06, *p* = .002, *d* = -0.15, 95% CI [-0.25, -0.05]. This is reminiscent of a contrast effect in which participants used their responses to Reditry as a contrastive benchmark, adjusting the sense of humor ideal down slightly.

Participants' estimates of the partner's sense of humor did not differ between the strong (*M* = 9.71, *SD* = 2.12) and weak (*M* = 9.71, *SD* = 2.10) functional preferences conditions, *t*(1637) = 0.08, *p* = .940, *d* = -0.004, 95% CI [-0.10, 0.09]. Participants' interest in joining a dating website with potential partners high in sense of humor was lower in the strong (*M* = 7.01, *SD* = 2.05) than the weak (*M* = 7.22, *SD* = 1.93) functional preferences condition, *t*(1636) = 2.13, *p* = .033, *d* = -0.11, 95% CI [-0.20, -0.01]. This last difference was again suggestive of a small contrast effect.

Mediation of partner Reditry on RQ2

We again ran an exploratory analysis examining whether the effect of condition on satisfaction was mediated by perceptions of partner's Reditry. As in Study 1, path a (i.e., condition on partner Reditry; β = 0.33, p <.001) and path b (i.e.,partner Reditry on satisfaction controlling for condition; β = 0.36, p <.001) were both significant, which suggests mediation according to the component approach of Yzerbyt et al. (2018). The direct effect (i.e., condition on satisfaction; β = -0.11, p <.001) was significant as well, but in the opposite of the predicted direction, suggesting a suppression effect. The indirect effect was again strong; β = 0.12, 95% CI [.10, .14].

Discussion

Study 2 replicated many of the findings from Study 1 using a high-powered sample and a stronger manipulation. The manipulation check again indicated that participants reported a higher ideal partner preference for Reditry in the strong than the weak condition. Critically, we again found support for the motivated projection account (RQ1) as well as the situation selection account (RQ4). These findings lend confidence to our conclusions that ideals *cause* people to perceive that their partner has a given trait, and ideals *cause* people to seek situations with prospective partners high on a given trait. Effects on our sense of humor control variable were small, and two of them were in the opposite direction, suggesting participants were contrasting their responses away from Reditry.

Although we powered the study in an attempt to detect a significant effect of the manipulation on relationship satisfaction (RQ2), no effect emerged. Nevertheless, because we preregistered our test and we used a high-powered sample, we attain some confidence that the effect size on satisfaction in this context is certainly likely to be quite small, if anything. The mediational pattern such that the manipulation boosted perceptions of Reditry, which was in turn associated with satisfaction, again emerged.

In contrast to Study 1, this study did find support for both the ideal partner-preferences account (RQ3a) and the moderated motivated perception account (RQ3b). In a way, these two findings are permutations of the same effect: Is the association between Reditry and satisfaction (regardless of which one is "IV" vs. "DV") different depending on condition?[14]

Finally, consistent with the perceiver effects account (RQ5), participants in the strong (vs. weak) condition perceived more Reditry in themselves, a friend, and a stranger, and less Reditry in a disliked other. Interestingly, this effect also seemed to reflect motivated reasoning, because the effect size tracks the evaluative nature of the targets. That is, the effect of the manipulation was highest on the self, followed closely by the friend judgment; people generally think very highly of themselves and their friends. The effect of the manipulation on the stranger was considerably lower (but still greater than zero). Finally, the effect on the disliked other was in the opposite direction; that is, when the manipulation caused participants to value Reditry highly, they thought the disliked other was *less* likely to have it. These findings are consistent with the perceiver effects, and suggest a motivated component to this causal account as well.

[14] For those who prefer this approach to conceptualizing this finding, the simple correlation between Reditry and satisfaction is r = .47 in the strong functional preference condition and r = .21 in the weak functional preference condition.

Study 3

In Study 3, we had two goals in mind. The first one was to rule out if experimenter expectations were driving the effects. For this purpose, we (a) added a funnel debriefing and the exclusion criteria to exclude all participants who had even the slightest suspicions of the study hypotheses or goals, and (b) added a measure of attitudes towards the study designed to probe for demand effects (Nichols & Maner, 2008).

The second goal was to test if the results replicated when using the real name of the trait. Studies 1-2 demonstrated that our design can manipulate the valence of a nonsense word, but to be sure that we are tapping into real-world partner preference dynamics, it would be helpful to show that our design can manipulate the valence of an attribute where there is some counterforce of reality. Thus, in Study 3 we told participants the trait was Youthfulness, instead of calling it Reditry; as long as our manipulation check is significant and a meaningful effect size, then the results for RQ1-5 will speak to the causal role of ideals in a setting with more real-world applicability. Finally, we added a measure to assess the connotative meaning of youthfulness (i.e., immature vs. energetic) to test whether the manipulation was causing a more positive conceptualization of the trait (Eastwick et al., 2011).

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The data, analysis code, materials and external (i.e., posted on OSF before data collection) preregistration can be found at https://osf.io/mptqk/?view_only=06b8f1fcbd0843c5b24e2b3353353a36. In this study, the manipulation check, RQs1-4, RQ5 (with the Bonferroni-Holm correction; Holm, 1979) and the new analyses were all preregistered.

Participants and Power

We estimated the sample size on InteractionPoweR (Finsaas et al., 2021; Baranger et at., 2022) aiming to replicate the interaction RQ3. We used the previous results from Study S2[15], Study 1 and Study 2 for the power analysis. Study S2 used the label Youthfulness instead of Reditry, and comparisons between Study 1 and Study S2 (which both used the Figures 9 and 10 manipulations) suggest that the effect sizes with "Youthfulness" are about one-third of the effect size obtained with "Reditry." In this study, we planned to use the more powerful manipulations from Study 2 and depicted in Figures 15 and 16.

Thus, to calculate the sample size, we needed four estimates in InteractionPoweR: *r* = .00 for the effect of condition on satisfaction, *r* = .27 for the effect of Youthfulness on satisfaction (taken from Study S2), *r* = .12 for the effect of condition on Youthfulness (1/3 of the size of the effect documented in Study 2) and *r* = .06 for the interaction (also 1/3 of the size documented in Study 2). The Shiny App calculated that we would need *N* = 1950 to reach 80% power for the interaction, and thus we aimed to collect *N* = 2500 to account for exclusions. If after exclusions and before running any analyses we have a sample size of less than 1900, we will calculate the exclusion rate ("e") and collect n = (1950-current n)/(1-e) additional participants.

Following this procedure, we ended up with *N* = 2475 who met the inclusion criteria used in Study 3 (i.e., between 18 and 35 years old, were primarily attracted to men, and in a romantic relationship with one male partner). We recorded *a priori* four exclusion criteria. The first three

[15] This study is in the Supplemental Materials and is similar to Study 1, except that we use the label "Youthfulness" instead of "Reditry".

were identical to Study 2. Besides these, we asked participants open-ended funnel debriefing questions about the goal of the study, and we set up to exclude participants who guessed that the Date-Fest task was designed to make them like Youthfulness more or less[16]. In total, *n* = 143 participants were excluded (*n* = 76 failed the "other" attention check, *n* = 2 failed the Winograd-like schema check, *n* = 0 failed the initials check and *n* = 17 showed suspicion of the study goal)[17].

Thus, after exclusions our final sample was *N* = 2332 (2169 women, 109 men, and 54 people who chose another option; Mage = 27.4, SD = 4.3; 63.9% White, 7.3% Black, 9.4% Asian or Pacific Islander, 10.0% Hispanic or Latino, 7.7% Biracial or Multiracial, 0.5% Native American and 1.2% other). Of these, *n* = 1160 were randomly assigned to the strong functional preference condition and *n* = 1172 were assigned to the weak functional preference condition, and they had been in a relationship of 63.9 months on average.

Procedure and Materials

The procedure and measures in this study were identical to Study 2, except for two changes. First, instead of using the label "Reditry", we told participants the trait was "Youthfulness" (which was true). Second, we added two more measures, as described below.

Trait meaning. In order to assess how participants interpreted the meaning of Youthfulness, we constructed a measure with three items based on Eastwick et al. (2011). We asked participants: "What do you think the word 'Youthfulness' means when you apply it to

[16] To be conservative, we excluded participants who reported even the slightest suspicion of the goal of the study. Participants guessed a wide variety of study goals (e.g., "[figuring out] what people look for in a romantic partner", "if people would like a dating app based on youthfulness"), but only 17 guessed it right in total or partially (e.g., "if the dating game influenced our relationship choices").

[17] We ran Study 3 on Prolific rather than MTurk, and exclusion rates were lower.

your current romantic partner [initials]?”. The three items ranged from -4 (immature, childish and juvenile) to 4 (energetic, enthusiastic and active), α = 85.

Attitudes towards the study. In order to test if participants were conforming to experimenter hypotheses, we adapted a scale from Nichols & Maner (2008). The four items (“How happy are you to do this study?”, “How important is it for you to be doing this study? “, ““How interested are you in the outcome of the study?”, “Would you recommend this study to a friend of yours?”; from 1 = not at all to 9 = very much) were averaged to create a scale of attitudes towards the study (α = .81). Nichols and Maner (2008) found that participants who scored highly on this scale were especially likely to conform to the study hypotheses when the hypotheses were obvious. Should a similar effect emerge in the current study—in this case, a positive interaction of this scale and study condition (implying that participants who liked the study were especially likely to provide data that fit the hypotheses)—it would suggest that our hypotheses might be exceptionally transparent to participants.

Results

Preference for Youthfulness (Manipulation Check)

As in the previous studies, participants indicated a stronger preference for Youthfulness in the strong (*M* = 5.87, *SD* = 1.59) than the weak (*M* =5.30, *SD* = 1.58) functional preferences condition, *t*(2330) = 8.72, *p* < .001, *d* = 0.36, 95% CI [0.28, 0.44]). In other words, the manipulation caused participants to boost their ratings of the importance of Youthfulness in an ideal romantic partner. As expected, because we told participants the trait was youthfulness, the magnitude of this manipulation was smaller (approximately 60% of Study 2). Nevertheless, the manipulation worked for this real-world trait.

Research Questions

RQ1: Estimate of Youthfulness in a current partner. Consistent with the motivated perception account, participants believed that their current partners had more Youthfulness in the strong (*M* = 8.61, *SD* = 2.11) than the weak (*M* = 8.09, *SD* = 2.17) functional preferences condition, *t*(2330) = 5.84, *p* < .001, *d* = 0.24, 95% CI [0.16, 0.32]). Similar to Studies 1 and 2, when the experimental manipulation caused participants to feel more positively about Youthfulness in an ideal romantic partner, they inferred that their partners had higher Youthfulness.

RQ2: Relationship satisfaction with a current partner. Participants' reports of their relationship satisfaction with their current partner was slightly higher in the strong (*M* = 7.78, *SD* = 1.46) than the weak (*M* = 7.67, *SD* = 1.55) functional preferences conditions. However, as in Study 1, this difference was only marginally significant, *t*(2330) = 1.85, *p* = .065, *d* = 0.08, 95% CI [-0.005, 0.16). We cannot conclude that ideal partner preferences causally boost downstream judgments of relationship satisfaction.

RQ3a: Ideal partner preference-matching (trait weighting). To test the preference-matching account, as in Study 2, we conducted a regression using condition (coded -1 = weak, 1 = strong), partner Youthfulness (standardized), and their interaction to predict relationship satisfaction (also standardized). Results are depicted in Figure 21. The main effect of condition was not significant, *β* = 0.01, *t*(2328) = 0.32, *p* = .746, and the main effect of Youthfulness was significant, *β* = 0.26, *t*(2328) = 13.11, *p* < .001. Importantly, the condition × Youthfulness interaction was not significant, *β* = 0.02, *t*(2328) = 0.91, *p* = .362. That is, in contrast to Study 2, the matching effect did not emerge: When participants were experimentally induced to feel more (vs. less) positively about Youthfulness, the extent to which they felt their partner had Youthfulness was not a stronger predictor of their relationship satisfaction.

RQ3b: Moderated motivated perception. To test the moderated motivated perception account, as in Study 2, we conducted a regression using condition (coded -1 = weak, 1 = strong), satisfaction (standardized), and their interaction to predict partner Youthfulness (also standardized). Results are depicted in Figure 22. The main effect of condition was significant as in the basic t-test in RQ1, *β* = 0.11, *t*(2328) = 5.54, *p* < .001, and the main effect of satisfaction was significant, *β* = 0.26, *t*(2328) = 13.16, *p* < .001. Importantly, in parallel with RQ3a and in contrast to Study 2, we did not find a significant condition × satisfaction interaction, *β* = 0.03, *t*(2328) = 1.42 , *p* = .157.

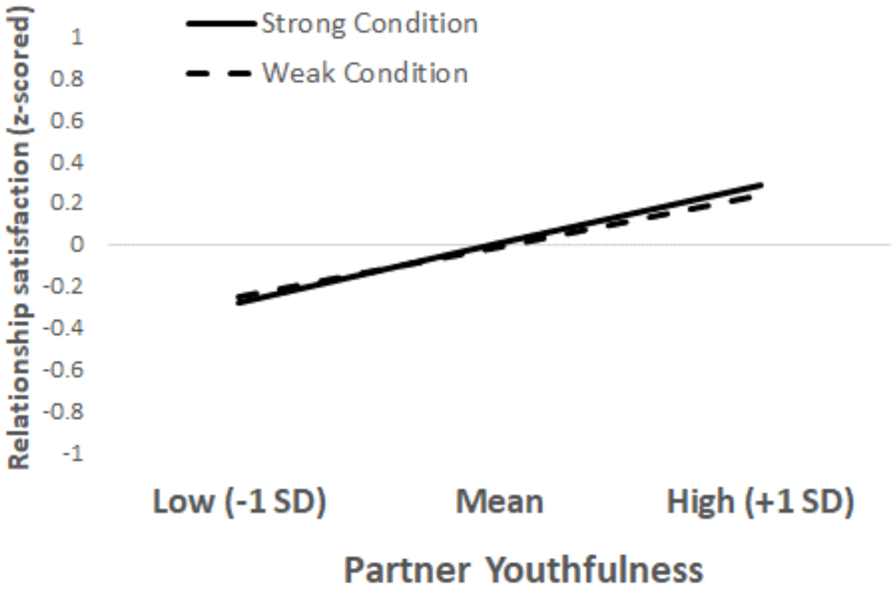


Figure 21. Regression results for RQ3a. The slope of Youthfulness predicting satisfaction did not differ by condition, which does not support the causal account for preference-matching.

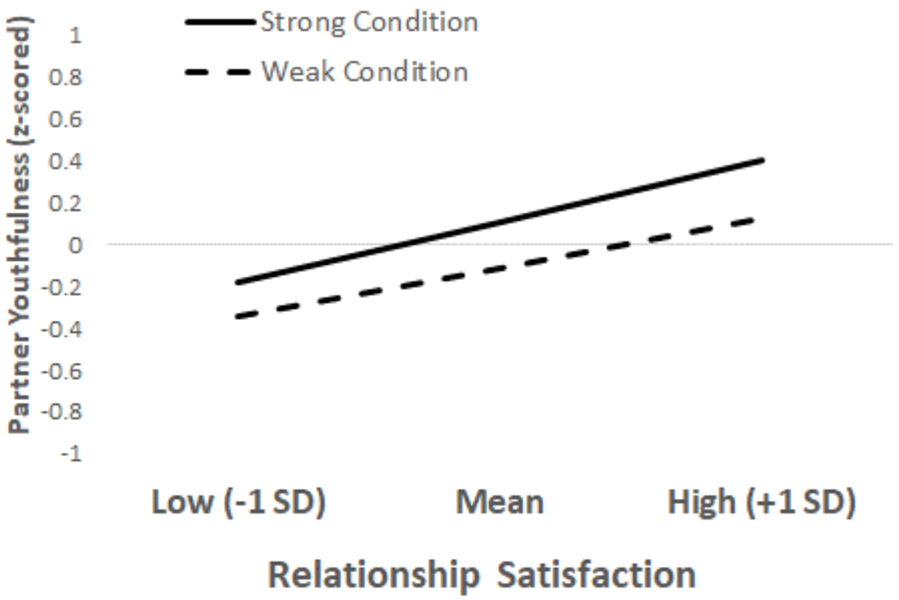


Figure 22. Regression results for RQ3b. The slope of satisfaction predicting Youthfulness did not differ by condition, which does not support the moderated motivated perception account.

RQ4: Situation selection. Participants reported greater interest in joining a website with potential partners high in Youthfulness in the strong (*M* = 4.67, *SD* = 2.33) than the weak (*M* = 4.04, *SD* = 2.21) functional preferences condition, *t*(2329) = 6.70, *p* < .001, *d* = 0.28, 95% CI [0.20, 0.36]. That is, once again, the manipulation boosted participants' interest in selecting into a situation containing partners high in Youthfulness.

Results for the manipulation check and RQ1-4 are depicted in Figure 23.

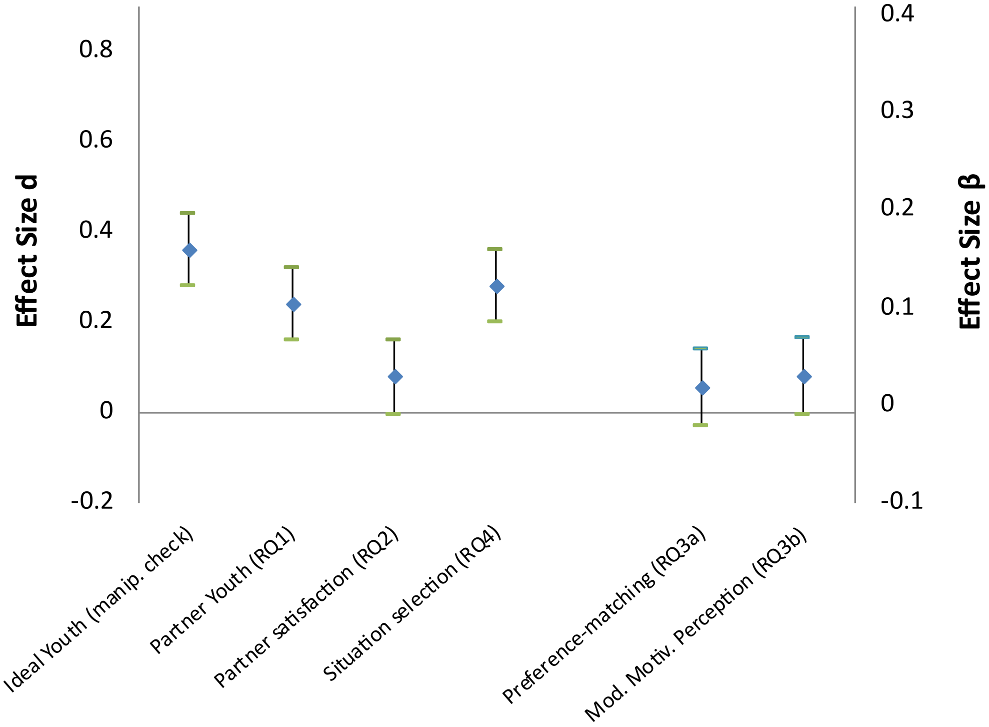


Figure 23. Effect sizes for Study 3. Left (4) effect sizes are *d*s; right (2) effect sizes are *β*s. Axes are scaled to be equivalent, $\beta = d \div \sqrt{(d^2 + 4)}$. Bars are 95% CIs.

RQ5: Perceiver effects. Did the manipulation also cause participants to boost the extent to which they thought *other* people possessed Youthfulness, besides their romantic partner? Figure 25 illustrates the effect sizes of the strong vs. weak condition on partner Youthfulness (RQ1), as well as ratings of Youthfulness in the self, a friend, a stranger, and a disliked other. Ratings of Youthfulness in the self were higher in the strong (M = 8.41, SD = 2.11) than the weak (M = 7.95, SD = 2.19) functional preferences condition, t(2330) = 5.16, p < .001, d = 0.21, 95% CI [0.13, 0.30]; ratings of Youthfulness in a friend were higher in the strong (M = 8.46, SD = 2.37) than the weak (M = 8.11, SD = 2.47) functional preferences condition, t(2329) = 3.47, p < .001, d = 0.14, 95% CI [0.06, 0.23]; ratings of Youthfulness in a stranger did not differ in the strong (M = 6.31, SD = 2.36) and the weak (M = 6.28, SD = 2.40) functional preferences condition, *t*(2330) = 0.36, *p* = .718, *d* = 0.01, 95% CI [-0.07, 0.10]; and ratings of Youthfulness in a disliked other were once again *lower* in the strong (*M* = 5.70, *SD* = 3.42) than the weak (*M* = 6.04, *SD* = 3.69) functional preferences condition, *t*(2328) = -2.35, *p* = .019, *d* = -0.10, 95% CI [-0.18, -0.02]. The three significant differences remained significant after a Bonferroni-Holm (Holm, 1979) correction, as preregistered.

In short, the manipulation affected participants' ratings of the Youthfulness of all targets except for the stranger. The direction and magnitude of the effects was once again consistent with a motivational account.

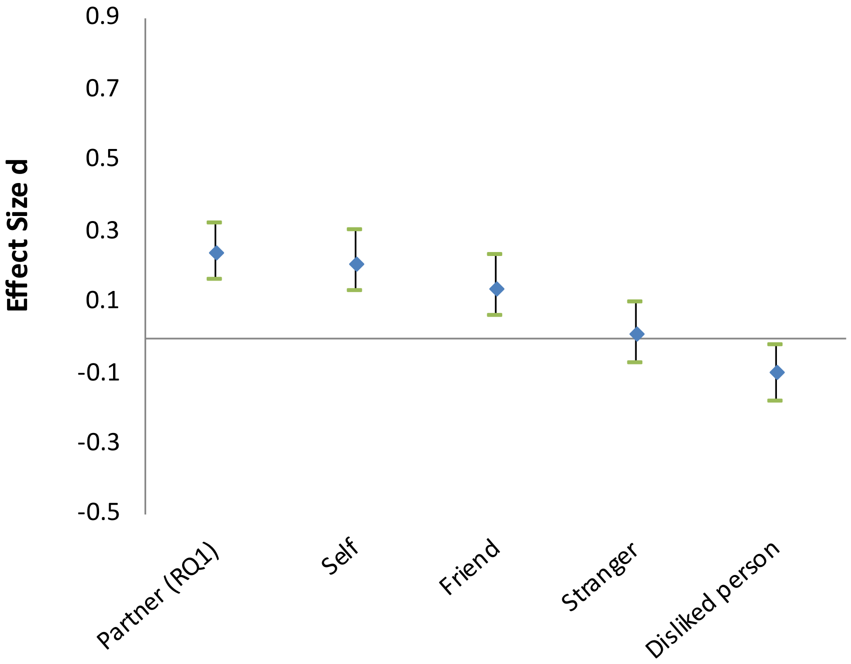


Figure 24. Effect sizes for RQ5, Study 3. Effect sizes (*d*) illustrate the effect of the manipulation on participants' ratings of Youthfulness of five different targets. Bars are 95% CIs.

Sense of Humor (Control) Questions

Again, we tested whether condition affected sense of humor. Participants' ideals for sense of humor did not significantly differ in the strong (*M* =8.10, *SD* = 1.00) than the weak (*M* = 8.12, *SD* = 0.93) functional preferences condition, *t*(2330) = 0.61, *p* = .545, *d* = -0.03, 95% CI [-0.11, -0.06].

Participants' estimates of the partner's sense of humor were higher in the strong (*M* = 10.21, *SD* = 1.76) than the weak (*M* = 9.89, *SD* = 1.93) functional preferences conditions, *t*(2330) = 4.19, *p* < .001, *d* = 0.17, 95% CI [0.09, 0.25]. Participants' interest in joining a dating website with potential partners high in sense of humor did not significantly differ in the strong (*M* = 6.89, *SD* = 2.12) and the weak (*M* = 6.97, *SD* = 1.95) functional preferences condition, *t*(2329) = 0.97, *p* = .330, *d* = -0.04, 95% CI [-0.12, 0.04].

New Analyses

Trait meaning. Participants' interpretation of the meaning of the trait "Youthfulness" as embodied by their partner was more positive in the strong (*M* = 7.20, *SD* = 1.49) than the weak (*M* =6.89, *SD* = 1.64) functional preferences conditions, *t*(2330) = 4.78, *p* < .001, *d* = 0.20, 95% CI [0.12, 0.28]. This finding suggests that the manipulation caused participants to believe that their partner exhibited an energetic, enthusiastic, and active (rather than an immature, childish, and juvenile) form of youthfulness. That is, consistent with a motivated reasoning account, participants who were induced to like the trait "youthfulness" seemed to believe that their partner was "the good" rather than "the bad" version of youthful.

Attitudes towards the study. We tested if attitudes towards the study acted as a moderator between condition and the effects RQ1-5. We did *not* find a significant condition × study liking interaction when predicting ideals for Youthfulness, *β* = 0.05, *t*(2327) = 0.25 , *p* = .805; partner Youthfulness, *β* = 0.02, *t*(2327) = 0.78 , *p* = .434; relationship satisfaction, *β* = 0.01, *t*(2327) = 0.52, *p* = .605; interest on the website with partners high on Youthfulness, *β* = 0.03, *t*(2326) = 1.69 , *p* = .090; estimates of Youthfulness in the self, *β* = 0.02, *t*(2327) = 0.84 , *p* = .400; in a disliked other, *β* = -0.02, *t*(2325) = -1.08 , *p* = .282; in a friend, *β* = -0.03, *t*(2326) = -1.65, *p* = .099; nor in a stranger, *β* = 0.002, *t*(2327) = 0.07, *p* = .942.

Study liking did significantly moderate RQ3a (i.e., the Condition × Study Liking × Partner Youthfulness 3-way interaction was significant), *β* = -0.07, *t*(2323) = -3.39 , *p* < .001, and RQ3b (i.e., the Condition × Study Liking × Satisfaction 3-way interaction was significant), *β* = -0.07, *t*(2323) = -3.70 , *p* < .001. But note that both of these interactions were negative, which is the opposite of the predicted direction (i.e., participants conformed to the study hypotheses to the extent that they *disliked* the study). We do not interpret these findings further, as they do not seem likely to be related to demand effects, and they seem likely to be spurious.

Mediation of partner Youthfulness on RQ2. We tested whether the effect of condition on satisfaction was mediated by partner's Youthfulness. Again, the Yzerbyt (2018) component approach suggested mediation: The results showed that path a (i.e., condition on partner Youthfulness; *β* = 0.12, p <.001) and path b (i.e.,partner Youthfulness on satisfaction controlling for condition; *β* = 0.26, p <.001) were both significant. The direct effect (i.e., condition on satisfaction; *β* = 0.01, p =.744) was not significant, and the indirect effect (*β* = 0.03, 95%CI [.02, .04]) accounted for most of the total effect.

Discussion

Study 3 replicated most of the findings from Study 2 using the name of the trait (youthfulness) instead of Reditry. Once again, we found support for the motivated projection (RQ1), situation selection (RQ4), perceiver effects (RQ5) and did not find support for the effect of condition on relationship satisfaction (RQ2). Critically, and contrary to Study 2, we did *not* find support for either the ideal partner-preferences account (RQ3a) nor the moderated motivated perception account (RQ3b). We also did not find evidence that the manipulation affected perceptions of a stranger, which suggests a limit on the perceiver effects account. Finally, we also tested if the effect of condition on satisfaction (RQ2) was mediated by partner's Youthfulness and found a pattern of total mediation.

Furthermore, participants in the strong functional preference condition interpreted the trait Youthfulness to have a more positive meaning in the strong than in the weak preference condition. We also wanted to rule out experimenter effects by checking if attitudes towards the study moderated the findings above. We did not find significant positive moderation for any of the effects, suggesting that our study was not so transparent that participants could provide results consistent with our hypotheses if they liked the task (Nichols & Maner, 2008).

General Discussion

The ideal partner preferences literature is enormous, dating back to the 1940s (Hill, 1945). Despite the importance of causal reasoning in theoretical approaches in this research area, none of these ideas have been tested using experimental approaches in prior research. This article removed a major roadblock by developing a replicable way to manipulate ideal partner preferences, which could be adapted by any researcher investigating causal consequences of ideals.

Four Causal Accounts of Ideals

The basic motivated projection account (RQ1) (Murray et al., 1996a, 1996b) received strong support: Ideal partner preferences caused people to perceive their current romantic partner highly on the relevant attribute. In other words, we found experimental support that ideals cause people to engage in "positive illusions" about their partners. Auxiliary findings in Studies 2 and

3 supported the motivated projection account by showing that this effect generalized to the self and a friend, and we even found a “negative illusions” effect for a disliked other. Although we could not document evidence that ideals caused participants to boost their satisfaction with their current partner (RQ2), a purported downstream consequence of positive illusions, we did find that partner perceptions on the relevant attribute mediated the effect of condition on satisfaction in all studies.

The situation selection account (RQ4) also received ample experimental support (e.g., da Silva Frost et. al., 2022). All three studies found that ideal partner preferences caused people to want to join a website featuring partners high on the relevant attribute, suggesting that ideals may cause people to shape their field-of-eligibles to have partners higher in the attribute.

The perceiver effects account (RQ5) received some degree of support. First and foremost, the manipulation boosted participants’ trait estimates for themselves and a friend in all cases, and it caused them to decrease their estimates for the extent to which a disliked other had the trait, too. This pattern of effect sizes suggested that closeness to the target was a key moderating variable, which is consistent with a motivated reasoning account. The target who had no relationship to the participant—the stranger—offers perhaps the most basic test of the perceiver effects account. These tests revealed partial support: Ideals caused participants’ trait estimates of a stranger to increase in Study 2, but not in Study 3.

Finally, the two moderational accounts (RQ3a and RQ3b) received support in Study 2 but did not receive support in Studies 1 or 3. We tentatively conclude that the causal role of ideals in these effects is likely quite small: It may be greater than zero, but it is likely quite hard to detect with conventional sample sizes.

RQ3a and RQ3b are two ways of examining the same interaction term: Did the manipulation boost satisfaction for participants who felt their partners had high Reditry/Youthfulness (ideal partner preference-matching), and did the manipulation boost judgments of Reditry/Youthfulness especially strongly for satisfied partners (moderated motivated perception)? When comparing these two accounts, it is important to note that ideals for the trait seemed more "movable" than relationship satisfaction (i.e., RQ1 vs. RQ2), so it might prove easier in other studies to find support for the moderated motivation perception account (where the trait moves) than the preference-matching account (where satisfaction moves). Also, these effect sizes were generally smaller than the motivated perception (RQ1) and situation selection (RQ4) effect sizes, but Study 2 suggests they may be greater than zero.

As for the Sense of Humor trait, which we did not attempt to manipulate, the effects were haphazard. In Study 1, there were no effects; in Study 2, we sometimes observed a contrast effect away from Reditry; and in Study 3, in one analysis we observed an assimilation effect in line with Youthfulness. In the Supplemental Materials, we present three more studies: two of which show no effect for Sense of Humor and one that shows a contrast effect. Overall, there was no noticeable pattern, and these differences are probably not meaningful.

Limitations and Strengths

Table 12

Assessment of Limitations

Dimension	Assessment
Internal validity	
Is the phenomenon diagnosed with experimental methods?	Yes
Is the phenomenon diagnosed with longitudinal methods?	No
Were the manipulations validated with manipulation checks, pretest data, or outcome data?	Yes, validated with manipulation checks in all studies.
What possible artifacts were ruled out?	We ruled out the possibility that the manipulation would boost the evaluation of other traits (e.g., sense of humor), and the possibility that the effects were driven by experimenter demand.
Statistical validity	
Was the statistical power at least 80%?	Yes, especially Studies 2 (powered to detect an effect size as small as *d* = .14) and 3 (powered to detect an interaction as small as *r* = .06).
Was the reliability of the dependent measure established in this publication or elsewhere in the literature?	Yes, reliabilities were reported in this article (and were strong throughout)
If covariates are used, have the researchers ensured they are not affected by the experimental manipulation before including them in comparisons across experimental groups?	No covariates were used.
Were the distributional properties of the variables examined and did the variables have sufficient variability to verify effects?	Yes; see descriptive tables in Supplemental Materials
Generalizability to different methods	

Dimension	Assessment
Were different experimental manipulations used?	We used different distributions of trials and some different face-stimuli in Study 1 (vs. Studies 2 and 3). However, we only used one paradigm, one type of stimuli (male faces from the Chicago Faces Database), and two traits (Youthfulness/ Reditry, sense of humor).
Generalizability to field settings	
Was the phenomenon assessed in a field setting?	No
Are the methods artificial?	Yes, the methods are artificial and constrained by the nature of the experimental manipulation. Thus, the effect sizes may not generalize to other contexts.
Generalizability to times and populations	
Are the results generalizable to different years and historic periods?	We do not know whether these findings would generalize. All samples were collected between 2019 and 2024.
Are the results generalizable across populations (e.g., different ages, cultures, or nationalities)?	We do not know whether these findings would generalize. All samples were American and online convenience samples, and all participants were attracted to men.
Theoretical limitations	
What are the main theoretical limitations?	Limitations include (a) we did not disentangle the two moderational accounts (i.e., moderated motivated projection and preference-matching); (b) we did not manipulate the purported underlying processes (e.g., motivation); and (c) we could not test alternative theoretical accounts that rely on thresholds, comparison points, or correlated patterns.

We chose to study youthfulness/babyfacedness because it is a real trait that can be perceived in faces, which was central to our manipulation. This trait was perhaps ideal in other ways, too. In a separate dataset of N = 1,266 participants (M_{age} = 25) who completed an online study[18], participants rated "youthful" as a $M = 7.7$, $SD = 2.3$ on a 1 (*not at all desirable*) to 11 (*highly desirable*) scale. In other words, "youthful" is considered on average to be a moderately desirable attribute, not unlike "extraverted/enthusiastic" (also $M = 7.7$, $SD = 2.3$ in this study), lower than an attribute like "attractive" ($M = 9.1$, $SD = 1.8$) and higher than an attribute like "religious" ($M = 4.7$, $SD = 3.4$).

Why was our choice fortuitous? A recent large international study ($N > 10{,}000$) found that moderately (but not highly) desirable attributes were especially likely to reveal ideal partner preference-matching effects (e.g., weighted matching effects were considerably larger for the moderately desirable trait "extraversion" than for highly desirable traits related to "warmth/trustworthiness" or "vitality/attractiveness"; Eastwick et al., 2024). Had we managed to craft a manipulation that changed participants' ideals for a trait like attractiveness, our lack of support for RQ3a and RQ3b would have been somewhat ambiguous in Study 3 (i.e., we would have chosen a context in which weighted ideal matching plays little role, causal or otherwise). Because we manipulated a trait that was moderately desirable, we put ourselves in the best position possible to find support for the preference-matching account.

This article also has several other strengths. Critically, it describes the tests of several influential theoretical accounts of ideals—the first experimental tests to establish causal relationships (Kenny, 1979; Falk & Heckman, 2009; Antonakis et. al., 2010). Experimental tests like these play a unique role in testing theory: they allow for precise control of the IVs, control of extraneous variables, reduction of error variance, and tests of moderational processes (Podsakoff & Podsakoff, 2019). In the current research, the experimental method allowed us to determine

[18] Please see Appendix B

the causal role ideals play in a myriad of relevant outcomes and further refine our understanding of the role of motivation and perception in romantic relationships.

References

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*(6), 1086-1120. https://doi.org/10.1016/j.leaqua.2010.10.010

Baranger, D. A. A., Finsaas, M. C., Goldstein, B. L., Vize, C. E., Lynam, D. R., Olino, T. M. (2023). “Tutorial: Power analyses for interaction effects in cross-sectional regressions.” *Advances in Methods and Practices in Psychological Science.*

Bowlby, J. (1988). *A secure base: Parent-child attachment and healthy human development.* New York: Basic Books.

Brandner, J. L., Brase, G. L., & Huxman, S. A. J. (2020). “Weighting” to find the right person: compensatory trait integrating versus alternative models to assess mate value. *Evolution and Human Behavior, 41*(4), 284-292.

Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences, 12*(1), 1-49. https://doi.org/10.1017/S0140525X00023992

Conroy-Beam, D., & Buss, D. M. (2016). Do mate preferences influence actual mating decisions? Evidence from computer simulations and three studies of mated couples. *Journal of Personality and Social Psychology, 111*(1), 53-66. https://doi.org/10.1037/pspi0000054

Conroy-Beam, D., Walter, K., & Duarte, K. (2022). What is a mate preference? Probing the computational format of mate preferences using couple simulation. *Evolution and Human Behavior, 43*(6), 510-526.

da Silva Frost, A., & Ledgerwood, A. (2020). Calibrate your confidence in research findings: A tutorial on improving research methods and practices. *Journal of Pacific Rim Psychology, 14*, e14. https://doi.org/10.1017/prp.2020.7

da Silva Frost, A., Wang, Y. A., Eastwick, P. W., & Ledgerwood, A. (2022). Summarized attribute preferences have unique antecedents and consequences. *Journal of Experimental Psychology: General*. Advance online publication. https://doi.org/10.1037/xge0001242

Eagly, A. H., Eastwick, P. W., & Johannesen-Schmidt, M. (2009). Possible selves in marital roles: The impact of the anticipated division of labor on the mate preferences of women and men. *Personality and Social Psychology Bulletin, 35*(4), 403-14. https://doi.org/10.1177/0146167208329696

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., Finkel, E. J., & Simpson, J. A. (2019). Best practices for testing the predictive validity of ideal partner preference-matching. Personality and Social Psychology Bulletin, 45, 167-181.

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., & Ledgerwood, A. (2019). How do people translate their experiences into abstract attribute preferences? *Journal of Experimental Social Psychology*, 85. https://doi.org/10.1016/j.jesp.2019.103837

Eastwick, P. W., Sparks, J., Finkel, E., Meza, E., Adamkovic, M., Adu, P., . . . Coles, N. (2024). A worldwide test of the predictive validity of ideal partner preference-matching. Stage 2 Registered Report Under Review, *Journal of Personality and Social Psychology*.

Erikson, E. H. (1959). *Identity and the life cycle*. New York: International Universities Press.

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science, 326*(5952), 535-538. https://doi.org/10.1126/science.1168244

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology, 87*(3), 293-311. https://doi.org/10.1037/0022-3514.87.3.293

Fazio, R. H., Pietri, E. S., Rocklage, M. D., & Shook, N. J. (2015). Positive versus negative valence: Asymmetries in attitude formation and generalization as fundamental individual differences. *Advances in experimental social psychology, 51,* 97-146. https://doi.org/10.1016/bs.aesp.2014.09.002

Finsaas, M. C., Baranger, D. A., Goldstein, B. L., Vize, C., Lynam, D., & Olino, T. M. (2021). InteractionPoweR Shiny App: Power analysis for interactions in linear regression. https://mfinsaas.shinyapps.io/InteractionPoweR/

Fletcher, G. J., Overall, N. C., & Campbell, L. (2020). Reconsidering "Best Practices" for Testing the Ideal Standards Model: A Response to Eastwick, Finkel, and Simpson (2018). *Personality & social psychology bulletin*, *46*(11), 1581–1595. https://doi.org/10.1177/0146167220910323

Fletcher, G. J., Simpson, J. A., Thomas, G., & Giles, L. (1999). Ideals in intimate relationships. *Journal of Personality and Social Psychology*, 76, 72-89. https://doi.org/10.1037/0022-3514.76.1.72

Fletcher, G. J., & Simpson, J. A. (2000). Ideal standards in close relationships: Their structure and functions. *Current Directions in Psychological Science, 9*(3), 102-105. https://doi.org/10.1111/1467-8721.00070

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*.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. *Perspectives on Psychological Science*, 15(5), 1243-1255.

Grosz, M. P., Ayaita A., Arslan R. C., Buecker, S., Ebert, T., Hunermund, P., Muller, S. R., Rieger, S., Zapko-Willmes, A., Rohrer, J. (2024). Natural Experiments: Missed Opportunities for Causal Inference in Psychology. *Advances in Methods and Practices in Psychological Science*, 7(1).

Hassebrauck, M. (1997). Cognitions of relationship quality: A prototype analysis of their structure and consequences. *Personal Relationships*, 4, 163-185.

Hill, R. (1945). Campus Values in Mate Selection. *Journal of Home Economics, 37*, 554-558.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70. https://doi.org/10.2307/4615733

Kenny, D. A. (1979). *Correlation and causality*. Hoboken: John Wiley & Sons.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis.* New York: Guilford Press.

Kille, D. R., Forest, A. L., & Wood, J. V. (2013). Tall, Dark, and Stable: Embodiment Motivates Mate Selection Preferences. *Psychological Science*, *24*(1), 112–114. https://doi.org/10.1177/0956797612457392

Kurzban, R., & Weeden, J. (2007). Do advertised preferences predict the behavior of speed daters? *Personal Relationships, 14*(4), 623–632. https://doi.org/10.1111/j.1475-6811.2007.00175.x

LaPrelle, J., Hoyle, R. H., Insko, C. A., & Bernthal, P. (1990). Interpersonal attraction and descriptions of the traits of others: Ideal similarity, self similarity, and liking. *Journal of Research in Personality, 24*(2), 216–240. https://doi.org/10.1016/0092-6566(90)90018-2

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.

Ledgerwood, A., & Shrout, P. E. (2011). The tradeoff between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology, 101,* 1174-1188.

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).

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*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Murray, S. L., Holmes, J. G., & Griffin, D. W. (1996a). The benefits of positive illusions: Idealization and the construction of satisfaction in close relationships. *Journal of Personality and Social Psychology, 70*(1), 79–98. https://doi.org/10.1037/0022-3514.70.1.79

Murray, S. L., Holmes, J. G., & Griffin, D. W. (1996b). The self-fulfilling nature of positive illusions in romantic relationships: Love is not blind, but prescient*.* *Journal of Personality and Social Psychology, 71*(6), 1155-1180. https://doi.org/10.1037/0022-3514.71.6.1155

Nelson, L. D., & Morrison, E. L. (2005). The symptoms of resource scarcity: Judgments of food and finances influence preferences for potential partners. *Psychological Science*, *16*(2), 167–173. https://doi.org/10.1111/j.0956-7976.2005.00798.x

Nichols, A. L., Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology, 135*(2), 151-165.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd edition). Cambridge University Press.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Casual inference in statistics: A primer.* Wiley.

Podsakoff, N. P., & Podsakoff, P. M. (2019). Methods for testing moderation: The past, present, and future. In R. E. Johnson (Ed.), *Cambridge Handbook of Testing and Measurement in Psychology* (pp. 341-360). Cambridge University Press.

Rau, R., Carlson, E., Back, M., Barranti, M., Gebauer, J., Human, L., Leising, D., & Nestler, S. (2021). What is the structure of perceiver effects? On the importance of global positivity and trait-specificity across personality domains and judgement contexts. *Journal of Personality and Social Psychology, 120*(3), 745-764. https://doi.org/10.1037/pspp0000278

Roskos-Ewoldsen, D. R., & Fazio, R. H. (1992). On the orienting value of attitudes: Attitude accessibility as a determinant of an object's attraction of visual attention. *Journal of Personality and Social Psychology, 63*(2), 198-211. https://doi.org/10.1037/0022-3514.63.2.198

Rusbult, C. E., & Buunk, B. P. (1993). Commitment processes in close relationships: An interdependence analysis. *Journal of Social and Personal Relationships, 10*, 175-204. http://dx.doi.org/10.1177/026540759301000202

Rusbult, C. E., Martz, J. M., & Agnew,C, R. (1998). The Investment Model Scale: Measuring commitment level, satisfaction level, quality of alternatives, and investment size. *Personal Relationships, 5*, 357-391.

Schellenberg, J. A. (1960). Homogamy in personal values and the “Field of eligibles”. *Social Forces*, *39*(2), 157-162.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Cengage Learning.

Simpson, J. A., Fletcher, G. J. O., & Campbell, L. (2001). The structure and function of ideal standards in close relationships. In Fletcher, G. J. O., & Clark, M. (Eds.), *The Blackwell handbook in social psychology: Interpersonal processes* (pp. 86-106). Oxford, UK: Blackwell.

Srivastava, S., Guglielmo, S., & Beer, J. S. (2010). Perceiving others' personalities: Examining the dimensionality, assumed similarity to the self, and stability of perceiver effects. *Journal of Personality and Social Psychology, 98*(3), 520–534. https://doi.org/10.1037/a0017057

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: *A social psychological perspective on mental health. Psychological Bulletin, 103*(2), 193-210. https://doi.org/10.1037/0033-2909.103.2.193

Thibaut, J. W., & Kelley, H. H. (1959). *The social psychology of groups*. New York: Wiley.

Yzerbyt, V., Muller, D., Batailler, C., Judd, C. M. (2018). New recommendations for testing indirects effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology: Attitudes and Social Cognition, 115*(6), 929-943.

Appendix A

Supplemental Materials for Chapter 2

Other Preregistered Analyses, Experiment 4

Following the preregistration, we broke down MCPRS into four factors. These factors were based on exploratory results of Experiment 3 data. Factor 1 was comprised of items 1, 6, 7, 10, 12, 13 and 14; factor 2 was comprised of items 2, 4, 16 and 17; factor 3 was comprised of items 3, 11 and 15; and factor 4 was comprised of items 5, 8 and 9. We then ran the same regression analyses as in the main article, except that we used the individual factors as moderators instead of the aggregated MCPRS scores. Thus, in one condition, we ran MCPRS factors moderating the relation between explicit category and IAT scores; in the other, we ran MCPRS factors moderating the relation between explicit category and explicit exemplar. The results were largely consistent with the aggregated MCPRS scores reported in the main article.

Figure S1.

Scree plot of MCPRS factors.

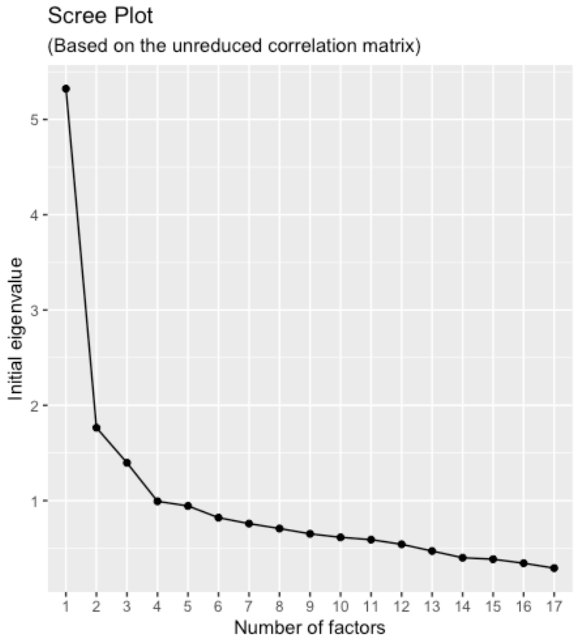


Figure S2. Correlations between MCPRS items.

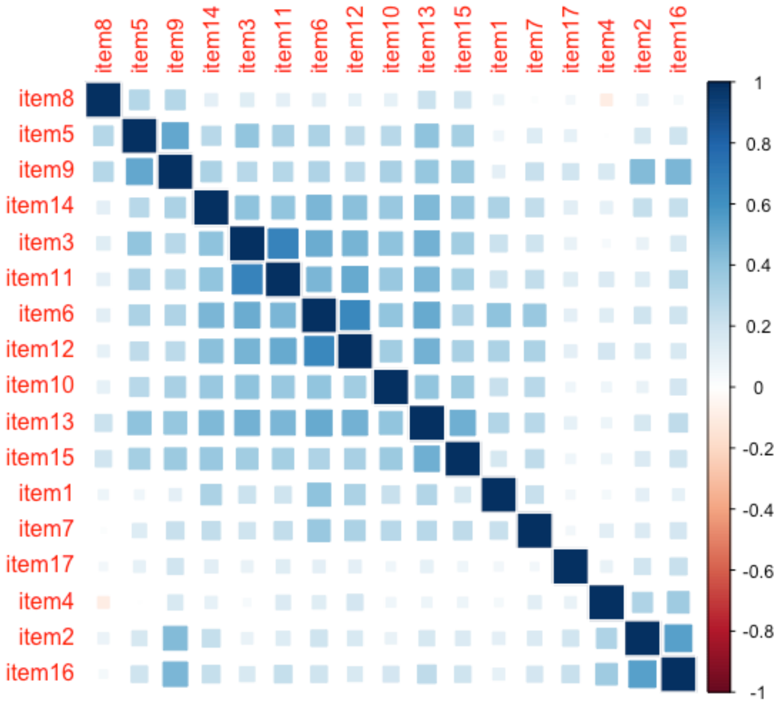


Table S1.

Descriptive statistics of and zero-order correlations between measures in the explicit exemplar condition.

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7
1. Exp. Cat.	539	-0.47	1.10	-						
2. MCPRS	539	0.54	0.90	-.14***	-					
3. Exp. Ex.	539	-0.24	0.84	.65***	-.19***	-				
4. MCPRS 1	539	1.12	1.00	-.09***	.87***	-.14	-			
5. MCPRS 2	539	-0.40	1.18	0.02	.64***	<.01	.36***	-		
6. MCPRS 3	539	0.54	1.46	-.19***	.81***	-.22***	.68***	.30***	-	
7. MCPRS 4	539	0.47	1.25	-.22***	.70***	-.27***	.45***	.33***	.50***	-

***p < .05, ***p < .001. Exp. Cat. = Explicit Category, Exp. Ex. = Explicit Exemplar.**

Table S2.

Results of multi-regression analyses using explicit exemplar evaluations and MCPRS factors to predict explicit category evaluations.

Model	Coefficients	*beta*	*SE*	*t*	*p*
1	Intercept	-0.48	0.04	-13.28	<.001
	MCPRS 1	<.001	0.04	0.008	.994
	Explicit Exemplar	0.71	0.04	19.26	<.001
	Interaction	-0.07	0.03	-2.16	.031
2	Intercept	-0.47	0.04	-13.05	<.001
	MCPRS 2	0.02	0.04	0.44	.660
	Explicit Exemplar	0.71	0.04	19.47	<.001
	Interaction	-0.01	0.03	-0.41	.680
3	Intercept	-0.48	0.04	-13.12	<.001
	MCPRS 3	-0.06	0.04	-1.59	.112
	Explicit Exemplar	0.70	0.04	18.76	<.001
	Interaction	-0.03	0.03	-1.21	.227
4	Intercept	-0.48	0.04	-12.96	<.001
	MCPRS 4	-0.06	0.04	-1.65	.099
	Explicit Exemplar	0.70	0.04	18.47	<.001
	Interaction	-0.02	0.03	-0.77	.444

All regressions have 535 degrees of freedom

Table S3.

Descriptive statistics of and zero-order correlations between measures in the IAT condition.

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7
1. Exp. Cat.	540	-0.42	1.07	-						
2. MCPRS	540	0.55	0.83	-.29***	-					
3. IAT	540	0.38	0.39	.24***	< .01	-				
4. MCPRS 1	540	1.13	0.94	-.24***	.86***	< .01	-			
5. MCPRS 2	540	-0.44	1.12	0.01	.56***	0.1*	.24***	-		
6. MCPRS 3	540	0.61	1.41	-.32***	.79***	-.05	.68***	.18***	-	
7. MCPRS 4	540	0.45	1.26	-.31***	.68***	-.08	.41***	.28***	.44***	-

***p < .05, ***p < .001. Exp. Cat. = Explicit Category**

Table S4.

Results of multi-regression analyses using IAT scores and MCPRS factors to predict explicit category evaluations.

Model	Coefficients	*beta*	*SE*	*t*	*p*
1	Intercept	-0.42	0.04	-9.60	<.001
	MCPRS 1	-0.26	0.04	-5.87	<.001
	IAT	0.25	0.04	5.81	<.001
	Interaction	-0.02	0.04	-0.34	.731
2	Intercept	-0.42	0.04	-9.43	<.001
	MCPRS 2	-0.008	0.05	-0.18	.857
	IAT	0.25	0.05	5.55	<.001
	Interaction	0.07	0.04	1.54	.125
3	Intercept	-0.42	0.04	-9.84	<.001
	MCPRS 3	-0.33	0.04	-7.64	<.001
	IAT	0.24	0.04	5.51	<.001
	Interaction	-0.03	0.04	-0.69	.493
4	Intercept	-0.42	0.04	-9.79	<.001
	MCPRS 4	-0.31	0.04	-7.32	<.001
	IAT	0.23	0.04	5.28	<.001
	Interaction	-0.03	0.04	-0.70	.482

All regressions have 536 degrees of freedom

Appendix B

Supplemental Materials for Chapter 3

Do People Experience Better Outcomes with High (vs. Low) Reditry Partners in the Strong vs. Weak Functional Preference Condition?

The goal of using the strong vs. weak functional preference conditions was to cause participants to have more positive experiences with high vs. low Reditry partners in the strong (vs. weak) condition. Of course, participants could choose their dates in all conditions, so the extent to which participants actually chose dates who were high (and rejected dates who were low) in Reditry is an empirical question. For all studies, we calculated the difference in Reditry in the dates participants accepted versus rejected separately for each condition (as in Eastwick, Smith, et al., 2019). For our manipulation to be successful, the difference in Reditry between the accepted and rejected dates should be larger in the strong vs. weak functional preference condition.

Study 1. The difference in Reditry between the accepted and rejected dates was larger in the strong functional preference condition (*M* = 33.82, *SD* = 14.98) than in the weak functional preferences condition (*M* = 24.86, *SD* = 15.52), *t*(401) = 5.88, *p* < .001, *d* = 0.59.

Study 2. Once again, the difference in Reditry between the accepted and rejected dates was larger in the strong functional preference condition (*M* = 41.86, *SD* = 11.48) than in the weak functional preferences condition (*M* =16.47, *SD* = 17.43), *t*(1637) = 34.96 *p* < .001, *d* = 1.73. The difference between this effect size and the Study 1 effect size indicates that we indeed made the manipulation quite a bit stronger in Study 2.

Study 3. As expected, the difference in Youthfulness between the accepted and rejected dates was larger in the strong functional preference condition (*M* = 39.85, *SD* = 9.81) than in the weak functional preferences condition (*M* =15.64, *SD* = 15.53), *t*(2313) = 44.78 *p* < .001, *d* = 1.86.

Other Preregistered Analyses, Study 1

Mediational model

We preregistered that we would test whether the ideal partner preference for Reditry statistically mediated the effect of condition on the partner Reditry estimate. Using the component approach recommended by Yzerbyt et al. (2018), we found evidence for mediation: The effect of condition on the mediator (ideal for Reditry) was significant, *β* = .19, *t*(401) = 3.80, *p* < .001, and the effect of the mediator (ideal for Reditry) on the dependent measure (partner Reditry) was significant in a regression that controlled for condition, *β* = .54, *t*(400) = 12.56, *p* < .001.

Moderation by Satisfaction of the Effectiveness of the Manipulation

We preregistered that we would test to see whether relationship satisfaction moderated the association between condition and participants' self-reported ideal partner preference for Reditry (manipulation check). The main effect of condition was significant, *β* = 0.18, *t*(399) = 3.61, *p* < .001, the main effect of satisfaction was significant, *β* = 0.11, *t*(399) = 2.21, *p* = .028, and the interaction was not significant, *β* = -0.06, *t*(399)= -1.30, *p* = .195. In other words, the manipulation boosted the manipulation check approximately the same for all participants, regardless of their level of relationship satisfaction.

Moderated Motivated Perception (Correlational)

We also preregistered that we would test to see whether relationship satisfaction moderated the association between participants' self-reported ideal partner preference for Reditry (rather than condition, as reported in the main manuscript as RQ3b) and the extent to which they thought their partner was high in Reditry. In this regression, all three variables were standardized. The main effect of ideals for Reditry was significant, $\beta = 0.49$, $t(399) = 12.80$, $p < .001$, the main effect of satisfaction was significant, $\beta = 0.37$, $t(399) = 9.53$, $p < .001$, and the interaction was significant, $\beta = 0.09$, $t(399)= 2.33$, $p = .021$. In other words, there was evidence for this "correlational test" of the moderated motivated perception account, but the analysis in the main manuscript did not find causal evidence for the moderated motivated perception account.

Other Preregistered Analyses, Study 2

Other Satisfaction Measures

We also asked participants to complete the relationship satisfaction measure about each target (self, α = .95; friend, α = .94; and disliked other, α = .94). The measures were identical to the original except that we replaced the target on each item (e.g., the original item "My relationship is close to ideal" was changed to "Our friendship is close to ideal" when the target was a friend and "My relationship with myself is close to ideal" when the target was the self).

Did the manipulation cause participants to boost their satisfaction with these three targets? Just as with satisfaction with one's romantic partner, the answer appeared to be "no." Satisfaction with oneself did not differ between the strong ($M = 6.14$, $SD = 1.92$) and the weak ($M = 6.04$, $SD = 2.01$) functional preferences condition, $t(1636) = 1.08$, $p = .279$, $d = 0.05$, 95% CI [-0.04, 0.15]; satisfaction with a friend did not differ between the strong ($M = 6.77$, $SD = 1.62$) and the weak ($M = 6.71$, $SD = 1.76$) functional preferences condition, $t(1622) = 0.73$, $p = .464$, $d = 0.04$, 95% CI [-0.06, 0.13]; and satisfaction with a disliked other did not differ between the strong ($M = 1.86$, $SD = 1.61$) and the weak ($M = 1.91$, $SD = 1.72$) functional preferences condition, $t(1631) = -0.71$, $p = .475$, $d = -0.04$, 95% CI [-0.13, 0.06].

Mediational model

As in Study 1, we preregistered that we would test whether the ideal partner preference for Reditry statistically mediated the effect of condition on the partner Reditry estimate. Using the Yzerbyt et al. (2018) approach, we found evidence for mediation: The effect of condition on the mediator (ideal for Reditry) was significant, $\beta = .36$, $t(1636) = 15.56$, $p < .001$, and the effect of the mediator (ideal for Reditry) on the dependent measure (partner Reditry) was significant in a regression that controlled for condition, $\beta = .47$, $t(1635) = 21.19$, $p < .001$.

Moderation by Satisfaction of the Effectiveness of the Manipulation

In this study, we preregistered that we would test to see whether relationship satisfaction moderated the effect of condition on the manipulation check (i.e., the ideal partner preference for Reditry). In this regression, the satisfaction and the ideal partner preference for Reditry measures were standardized, and condition was coded -1 = weak, 1 = strong. The main effect of condition was significant, $\beta = 0.36$, $t(1634) = 15.56$, $p < .001$, the main effect of satisfaction was not significant, $\beta = 0.04$, $t(1634) = 1.62$, $p = .106$, and their interaction was not significant, $\beta = 0.04$, $t(1634)= 1.84$, $p = .065$. In other words, the manipulation boosted the manipulation check approximately the same for all participants, regardless of their level of relationship satisfaction.

Study S1a

Study S1a aimed to establish a paradigm for manipulating ideal partner preferences. For this purpose, there were two between-subjects conditions: strong vs. weak functional preference for the focal trait “Reditry.” Specifically, in the strong functional preference condition, Reditry was strongly associated with the likelihood of going on enjoyable dates, and in the weak functional preference condition, the association was only modest. That is, participants’ experiences with higher levels of Reditry was more positive in the strong than the weak condition.

The study used a modified DateFest paradigm and an unfamiliar name for babyfacedness (Reditry) to recreate the experience of forming a preference in the first place and to circumvent participants’ existing beliefs about babyfacedness (da Silva Frost et al., 2022). The dependent measure—which serves as the manipulation check in Studies 1-3 in the main manuscript—was participants’ self-reported ideal partner preference (i.e., their “stated” or “summarized” preference) for Reditry.

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The data, analysis code, materials and internal (i.e., on a lab server) preregistration can be found at https://osf.io/mptqk/?view_only=06b8f1fcbd0843c5b24e2b3353353a36. In this study, we preregistered the analysis under “preference for Reditry” in the Results section below.

Participants and Power

Our final sample consisted of *N* = 831 (743 females, 83 males, and 5 people who chose another option; M_{age} = 28.3, *SD* = 4.3; 66.7% White, 10.6% Black, 8.1% Asian or Pacific Islander, 7.7% Hispanic or Latino, 4.6% Biracial or Multiracial, 1.2% Native American and 1.1% other). We used only male faces as stimuli because the average level of babyfacedness and the association between babyfacedness and attractiveness differ by sex (Ma et al., 2015). For this reason, we decided *a priori* to include only participants who reported being mostly attracted to men. Furthermore, we decided *a priori* to include only participants in the 18-35 age range, to match the apparent age of the Chicago Face Database stimuli. *N* = 569 of the participants (68.5%) reported being currently involved in a romantic relationship (relationship length *M* = 66 months).

Our final sample of 831 (after exclusions) came from a larger sample of 1,064 participants who completed the study online through the MTurk platform. We set and recorded *a priori* to target a sample size of 800, and we collected many more because we anticipated (based on prior experience) a high number of exclusions (~25%). We set and recorded the following *a priori* exclusion criteria: We would exclude participants who (1) failed to select "Other" as instructed in our attention check, and/or (2) provided a nonsensical response to a Winograd-like schema designed to filter out bots or inattentive participants. On total, *n* = 233 participants were excluded (*n* = 143 failed the "other" attention check, *n* = 110 failed the Winograd-like schema check, and *n* = 20 failed both).

Participants were randomly assigned to one of four between-subjects conditions in a 2 (functional preference for Reditry: strong vs. weak) x 2 (complexity: high vs. low) design. A sensitivity analysis indicated that our sample (*N* = 831, or ~208 per cell) provided 90% power to detect a simple effect of the functional preference manipulation (within each complexity condition) of *d* = .32 and 80% power for an effect of *d* = .28. For reference, the median effect size in social psychology has been estimated at *r* = .21, or approximately *d* = 0.43 (Richard, Bond & Stokes-Zoota, 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 men were able to proceed, according to the inclusion criteria.

Covariation detection task. Participants then played a game called DateFest (Eastwick, Smith, et al., 2019) in which the goal was to gain points by making rewarding dating decisions. Participants were told that they were going to a party where they would meet 24 party guests and that they must decide whether to go on a date with each one. Participants learned that some of the dates would be good experiences whereas others would be bad experiences, and they needed to figure out which guests would lead to good vs. bad experiences in order to gain points. Choosing to go on the dates that were a bad experience (which was true for 12 out of 24 guests) caused participants to lose 10 points, and choosing dates that were a good experience (the other 12 guests) caused them to gain 10 points.

Participants only lost or gained points when they chose to go on dates. If they chose not to go on a date with a guest, they neither gained nor lost any points. To ensure participants had access to the same amount of information regardless of how many dates they chose to go on, they learned whether they would have gained or lost points if they had gone on the date. Therefore, the task was designed to induce participants to form more positive evaluations when the traits were associated with gaining (vs. losing) points.

Manipulating functional preference strength. The manipulation was the same as Study 1 in the manuscript (see also Figures 9, 10 and 11 in the main text).

Manipulating complexity. Because we also wanted to investigate whether the number of traits present in the paradigm affects the strength of the functional preference manipulation, we included a second manipulation: *complexity*. In the low complexity conditions, participants had to track only the focal trait (Reditry), whereas in the high complexity conditions, participants had to track the focal trait and an additional control trait (sense of humor).

In the low complexity condition, participants saw each party guest's face and learned how much Reditry the guest had before making a "date" vs. "do not date" decision. In the high complexity condition, participants saw each face and learned about each guest's Reditry and sense of humor.

As mentioned above, Reditry was babyfacedness as measured in the CFD norming data (Ma et al., 2015); we selected 24 faces from this database to use as stimuli (babyfacedness *M* = 2.55, *SD* = 0.70, range = 1 to 7) and we algebraically transformed these values to range from 1 to 120. We did not manipulate the functional preference strength of sense of humor, and these values were not connected to features of the face. To ensure that sense of humor was equally likable across both the weak and strong Reditry functional preference conditions, the good dates always had an average sense of humor of 85 and the bad dates always had an average sense of humor of 53. Reditry and sense of humor levels were chosen so that the two traits did not correlate with each another (*r*s ranged from = -.01 to .02) within the set of liked targets and within the set of disliked targets; see Supplemental Materials for details on stimuli.

Ideal partner preference measures. After playing DateFest, participants responded to the following four questions, which comprised the ideal partner preference for Reditry dependent measure: “How important is Reditry to you in a romantic partner?”, “How much do you value Reditry in a romantic partner?,” “How desirable is Reditry to you in a romantic partner?,” and “To what extent does Reditry characterize your ideal romantic partner?” on scales from 1 (not at all) to 9 (extremely). In the high complexity condition, they responded to these same questions about sense of humor as well. These four items were highly reliable (α = .97 for Reditry and α = .94 for sense of humor) and were thus averaged to form a scale reflecting participants' summarized preference for Reditry and sense of humor.

Situation selection items. We included items assessing situation selection in all three studies (RQ4). To assess participants' interest in entering a situation with potential partners high in Reditry, participants read the following prompt: “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. How interested are you in the website that would only include partners high in Reditry?” They responded on a 9-point Likert-type scale (from 1 = *not at all interested* to 9 = *very interested*). Participants then completed an identical item about sense of humor.

Attention checks. We included two attention checks. The first one was a question in the demographics section that instructed participants to select “other” instead of their actual region of origin. Participants who did not follow the instructions were excluded from the analysis. The second attention check 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, Davis, & Morgenstern, 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., "unfortunately"), as coded by a researcher blind to the study results.

Results

Preference for Reditry

In this study, we were primarily interested in whether the functional preference manipulation would cause participants to self-report a stronger ideal partner preference for Reditry. A 2 (functional preference: strong vs. weak) x 2 (complexity: low vs. high) between-subjects ANOVA indicated that there was a main effect of the functional preference condition, $F(1, 827) = 74.61, p < .001$, *partial* $\eta^2 = .08$. As anticipated, participants reported more of a preference for Reditry in the strong ($M = 6.32$, $SD = 2.16$, $N = 421$) than the weak ($M = 4.98$, $SD = 2.37$, $N = 410$) condition, $d = 0.59$, 95% CI [0.45, 0.73]. There was also an unanticipated main effect of complexity, $F(1, 827) = 24.18, p < .001$, *partial* $\eta^2 = .03$, such that participants in the high complexity condition reported greater liking for Reditry ($M = 6.03$, $SD = 2.04$, $N = 428$) than participants in the low complexity ($M = 5.27$, $SD = 2.60$, $N = 403$) condition, $d = 0.33$. The functional preference × complexity interaction was not significant, $F(1, 827) = 2.41, p = .121$, *partial* $\eta^2 = .003$.

RQ4: Situation Selection

Participants reported greater interest in joining a website with potential partners high in Reditry in the strong ($M = 6.17$, $SD = 2.40$) than the weak ($M = 5.39$, $SD = 2.59$) functional preferences condition, $t(829) = 4.51$, $p < .001$, $d = 0.31$, 95% CI [0.18, 0.45]. In other words, consistent with the situational selection account, the experimental manipulation boosted participants' interest in selecting into a situation containing partners high in Reditry.

Difference in accepted dates

As expected, the difference in Reditry between the accepted and rejected dates was larger in the strong functional preference condition ($M = 34.71$, $SD = 12.47$) than in the weak functional preference condition ($M = 24.50$, $SD = 15.53$), $F(1, 823) = 109.06$, $p < .001$, *partial* $\eta^2 = .12$. This main effect was not moderated by the complexity condition: functional preference strength × complexity interaction $F(1, 823) = 1.11$, $p = .293$, *partial* $\eta^2 = .001$.

Sense of Humor (Control) Questions

We also tested whether condition affected sense of humor, a trait that was not ostensibly connected to the manipulation. We conducted an independent samples *t*-test comparing the strong to weak functional preferences conditions within the high complexity condition. (The low complexity conditions did not enter the analysis since there were no sense of humor values in that condition, and we did not assess these participants' preferences for sense of humor.) Participants actually indicated a weaker preference for sense of humor in the strong functional preference ($M = 7.68$, $SD = 1.25$) than the weak functional preference ($M =7.93$, $SD = 1.06$) condition, $t(426) = -2.22$, $p = 0.027$, $d = -0.21$, 95% CI [-0.03, -0.41]. These results suggest that participants might have been induced to contrast their sense of humor preferences away from their Reditry preferences. Participants also reported less interest in joining a website with potential partners high in sense of humor in the strong ($M = 6.74$, $SD = 2.04$) than the weak ($M = 7.25$, $SD = 1.82$) functional preferences condition, $t(426) = -2.72$, $p = .007$, $d = -0.26$, 95% CI [-

0.45, -0.07]. That is, the effect went in the opposite direction of the Reditry website DV and is consistent with the contrast effect for the sense of humor preference.

Discussion

Study S1a used the modified DateFest paradigm to manipulate ideal partner preferences for a novel attribute. In the strong functional preference condition, Reditry—which was actually babyfacedness—was strongly associated with the rewards that participants experienced on their dates. In the weak functional preference condition, Reditry was only modestly connected to the rewards that participants experienced.

Post-game, participants in the strong (vs. weak) functional preference condition reported a stronger preference for Reditry in a romantic partner, with a moderate-to-strong effect size. The result did not differ depending on whether participants had to track one or two traits (i.e., the complexity manipulation). In line with the situation selection account (RQ4), this manipulation also caused participants to report more interest in joining a website featuring potential partners high in Reditry.

Study S1b

Study S1b was identical to Study S1a, except that participants saw the real name of the trait, “youthfulness”, instead of the made-up label “Reditry”. The other two differences were that we counterbalanced the order of the sense of humor and youthfulness DVs, and we did not investigate complexity (i.e., all participants learned about youthfulness and sense of humor on each trial, as in Study 1).

Method

Participants and power. Our final sample consisted of N = 829 (727 females, 94 males, and 8 people who chose another option; M_{age} = 28.1, SD = 4.5; 61.5% White, 13.6% Black, 7.8%

Asian or Pacific Islander, 7.7% Hispanic or Latino, 5.2% Biracial or Multiracial, 2.3% Native American and 1.7% other). As in Study S1a, we decided *a priori* to include only participants who reported being mostly attracted to men and in the 18-35 age range.

Our final sample of 829 (after exclusions) came from a larger sample of 1,030 participants who completed the study online through the MTurk platform. Because we expected a smaller effect size due to using the real name of the trait, we decided to double the cell size of Study S1: That is, we aimed for at least 400 participants per cell. Therefore, we set and recorded *a priori* to target a sample size of 800, and we collected many more because we anticipated (based on prior experience) a high number of exclusions (~25%). In total, n = 201 participants were excluded (n = 154 failed the "other" attention check, n = 85 failed the Winograd-like schema check, n = 38 failed both). A sensitivity analysis indicated that this sample size provided 90% power to detect a difference a main effect of d = .23 and 80% to detect an effect of d = .20.

Procedure. The procedure was exactly the same as Study S1a, with three important differences: we used the real name of the trait, youthfulness, as opposed to Reditry; the order of the youthfulness and sense of humor summarized preference items was counterbalanced; and all party guests had both youthfulness and sense of humor values (i.e., all participants were in the high complexity conditions, as in the studies in the main manuscript).

Results

Preference for Youthfulness

A planned independent samples t-test examined whether participants reported greater liking for youthfulness in the strong than the weak functional preferences condition. As in Study S1a, participants indicated a higher ideal partner preference for youthfulness in the strong (M = 5.91, SD = 1.79) than the weak (M =5.49, SD = 1.71) condition, $t(827) = 3.47$, $p < .001$, $d = 0.24$, 95% CI [0.10, 0.38]. In other words, the manipulation caused participants to boost their ratings of the importance of youthfulness in an ideal romantic partner; this effect size was approximately 70% of the size of the effect in Study S1a.

RQ4: Situation Selection

We did not find evidence that participants reported greater interest in joining a website with potential partners high in youthfulness in the strong (*M* = 4.80, *SD* = 2.24) than the weak (*M* = 4.58, *SD* = 2.31) functional preferences condition, *t*(827) = 1.39, *p* = .164, *d* = 0.10, 95% CI [-0.04, 0.23]. However, it is perhaps worth noting that the effect size here was also approximately 70% of the size of the effect in Study S1a.

Preference for Sense of Humor

A planned independent samples t-test examined whether participants reported greater liking for sense of humor in the strong than the weak functional preferences condition. There was no evidence that participants' ideals for sense of humor differed between the strong (*M* = 7.94, *SD* = 1.04) and the weak (*M* = 7.87, *SD* = 1.08) condition, *t*(827) = 0.97, *p* = .332, *d* = 0.07, 95% CI [-0.07, 0.20]. In other words, there is no evidence the manipulation affected ratings of the importance of sense of humor in an ideal romantic partner. Similarly, we did not find evidence that participants reported greater interest in joining a website with potential partners high in sense of humor in the strong (*M* = 6.63, *SD* = 2.02) than the weak (*M* = 6.69, *SD* = 1.99) functional preferences condition, *t*(827) = -0.47, *p* = .638, *d* = -0.03, 95% CI [-0.17, 0.10].

Repeated Measures ANOVA

As planned in the preregistration, we ran an ANOVA with condition x DV type (youthfulness vs. sense of humor) with repeated measures on the DV type factor. The effect of condition was significant *F*(1, 827) = 12.48, *p* < .001, ges = .007, suggesting there were differences between the two conditions collapsing across DV type. The main effect of DV type was also significant, *F*(1, 827) = 925.78, *p*< .001, ges = .367, suggesting there were differences between the youthfulness DV and sense of humor DV. The interaction was also significant, *F*(1, 827) = 5.87, *p* = .016, ges = .004, suggesting that the effect of condition on the dependent variable was different depending on the DV type.

Counterbalanced DV Order

As planned in the preregistration, we ran a 2 (condition: strong vs. weak functional preference) x 2 (counterbalanced order: youthfulness DV first or sense of humor DV first) between-subjects ANOVA on ideals for youthfulness. The effect of condition was significant, *F*(1, 825) = 12.02, *p* < .001, η^2 = .014, suggesting condition affected ideals for youthfulness. The effect of order was not significant, *F*(1, 825) = .01, *p* = .926, η^2 < .001, suggesting there were no differences in ideals depending on DV order. The interaction was not significant*, F*(1, 825) = 0.09, *p* = .759, η^2 < .001, suggesting the effect of condition on ideals for youthfulness did not depend on the order in which participants saw the DV. We did not run simple effects since the interaction was not significant.

We ran the same model with ideals for sense of humor as DV. The effect of condition was not significant, *F*(1, 825) = 0.95, p = .331, η^2 = .001, suggesting condition did not affect ideals for sense of humor. The effect of order was significant, *F*(1, 825) = 7.32, p = .007, η^2 = .009, suggesting there were differences in ideals depending on DV order. The interaction was not significant, *F*(1, 825) = 0.026*, p* = .871, η^2 < .001, suggesting the effect of condition on ideals for sense of humor did not depend on the order in which participants saw the DV. We did not run simple effects since the interaction was not significant.

Discussion

In this study, we used the label “youthfulness” instead of “Reditry”. The results indicated that the manipulation still worked, but the effect sizes were smaller, potentially because it was harder to shift participants’ preconceived notions of how much they like the trait.

Study S2

Study S2 was identical to Study 1, except that we used the real name of the trait “youthfulness” instead of the made-up label “Reditry”. Besides the label, we also had an item to assess participants’ estimate of youthfulness in a platonic friend, and a measure for participants to estimate youthfulness in new faces. RQ2, RQ3a, RQ3b, and RQ4 were not preregistered.

Method

Participants and power. Because we expected a small effect but were unsure of the size, we planned and preregistered a sequential analysis with total sample of 2500 after exclusions, divided in batches of 4 with 625 each (for more on sequential analyses, see Lakens, 2014; da Silva Frost & Ledgerwood, 2020). That is, we would pause data collection once we had at least 625 participants who pass all exclusion checks, and we would stop all data collection if the *p* value for the main analysis (effect of condition on estimation of partner’s youthfulness) is less than .013. If the *p* value is higher than this number, we would proceed to collect another 625 participants, and we would stop all data collection if the *p* value for the “main analysis” is less than .016. If the *p* value is higher than this number, we would proceed to collect another 625 participants, and we would stop all data collection if the *p* value for the “main analysis” is less than .020. If the *p* value is higher than this number, we would proceed to collect another 625 participants, and we would conclude support for the “main analysis” only if the *p* value is less than .025.

Because the *p* value did not make the adjusted cut-off at any point, we proceeded to collect data up until the last planned batch. Our final sample consisted of *N* = 2543 (2209 females, 316 males, and 18 people who chose another option; M_{age} = 27.2, *SD* = 4.7; 62.9% White, 9.4% Hispanic or Latino, 8.5% Black, 8.2% Asian or Pacific Islander, 5.5% Biracial or Multiracial, 2.9% Native American and 2.6% other).

Our final sample of 2543 (after exclusions) came from a larger sample of 3515 participants who completed the study online through the MTurk platform. On total, *n* = 971 participants were excluded (*n* = 588 failed the "other" attention check, *n* = 570 failed the Winograd-like schema check, and *n* = 187 failed both).

Procedure

The procedure was the same as Study 2, using youthfulness instead of Reditry and with the added measures of youthfulness estimation in a platonic friend (i.e., "If you had to guess, how much Youthfulness do you think your closest platonic friend has?"). We also added a measure to estimate youthfulness in stranger faces.

Estimate of youthfulness in faces. After DateFest and the main DVs, we presented participants with 12 additional male faces from the CFD, in random order. For each face, we asked them to estimate the amount of youthfulness from 1 (very little youthfulness) to 12 (a lot of youthfulness).

Results

Preference for Youthfulness (Manipulation Check)

A planned independent samples t-test examined whether participants reported greater liking for youthfulness in the strong than the weak functional preferences condition. As in Study S1b, participants indicated a higher ideal partner preference for Youthfulness in the strong (*M* =

5.97, *SD* = 1.77) than the weak (*M* =5.73, *SD* = 1.77) condition, *t*(2541) = 3.44, *p* < .001, *d* = 0.14, 95% CI [0.06, 0.21]. In other words, the manipulation caused participants to boost their ratings of the importance of youthfulness in an ideal romantic partner.

Research Questions

RQ1: Estimate of youthfulness in a current partner. We tested whether participants believed that their current partners had more youthfulness in the strong (*M* = 8.37, *SD* = 2.11) than the weak (*M* = 8.20, *SD* = 2.23) functional preferences condition, *t*(2541) = 1.98, *p* = .048, *d* = 0.08, 95% CI [0.0007, 0.16]. Because we ran a sequential analysis (see above in the "Participants and power" section), this result was not significant. Thus, we cannot conclude that the experimental manipulation caused participants to infer that their partners had higher youthfulness.

RQ2: Relationship satisfaction with a current partner. We tested whether participants' reports of their relationship satisfaction differed between the strong (*M* = 7.49, *SD* = 1.50) from the weak (*M* = 7.56, *SD* = 1.42) functional preferences condition, but this difference was not significant, *t*(2541) = -1.19, *p* = .236, *d* = -0.05, 95% CI [-0.12, 0.03]. We cannot conclude that ideal partner preferences causally boosted downstream judgments of relationship satisfaction.

RQ3a: Ideal partner preference-matching. To test the preference-matching account, we conducted a regression using condition (coded -1 = weak, 1 = strong), partner youthfulness (standardized), and their interaction to predict relationship satisfaction (also standardized). The main effect of condition was not significant, β = -0.03, *t*(2539) = -1.79, *p* = .074, and the main effect of partner youthfulness was significant, β = 0.27, *t*(2539) = 14.15, *p* < .001. Critically, as in Study 1, the interaction term was not significant, β = 0.004, *t*(2539) = 0.20, *p* = .846.

RQ3b: Moderated motivated perception. To test the moderated motivated perception account, we conducted a regression using condition (coded -1 = weak, 1 = strong), satisfaction (standardized), and their interaction to predict partner youthfulness (also standardized). The main effect of condition was significant, $\beta = 0.05$, $t(2539) = 2.39$, $p = .017$, and the main effect of satisfaction was significant, $\beta = 0.27$, $t(2539) = 14.22$, $p < .001$. As with RQ3a, the interaction was not significant, $\beta = -0.03$, $t(2539) = -1.36$, $p = .175$.

RQ4: Situation selection. Participants reported greater interest in joining a website with potential partners high in youthfulness in the strong ($M = 5.38$, $SD = 2.36$) than the weak ($M = 4.96$, $SD = 2.37$) functional preferences condition, $t(2541) = 4.47$, $p < .001$, $d = 0.18$, 95% CI [0.10, 0.26]. In other words, the experimental manipulation boosted participants' interest in selecting into a situation containing partners high in youthfulness.

RQ5: Perceiver effects. Participants did not report their close friend had more youthfulness in the strong ($M = 8.33$, $SD = 2.11$) than the weak ($M = 8.21$, $SD = 2.14$) functional preferences condition, $t(2541) = 1.41$, $p = .158$, $d = 0.06$, 95% CI [-0.02, 0.13].

Sense of Humor (Control) Questions

We also tested whether condition affected sense of humor, a trait that was unrelated to the manipulation. Participants' ideals for sense of humor did not differ across the strong ($M = 8.01$, $SD = 1.02$) vs. weak ($M = 8.01$, $SD = 1.06$) functional preference conditions, $t(2541) = 0.12$, $p = 0.907$, $d = 0.005$, 95% CI [-0.07, 0.08]. In other words, our manipulation did not affect sense of humor ideals in this study.

Not surprisingly, participants' estimates of the extent to which their partner has a sense of humor did not differ across the strong ($M = 9.47$, $SD = 2.05$) vs. weak ($M = 9.59$, $SD = 2.02$) functional preference conditions, $t(2541) = -1.43$, $p = .153$, $d = -0.06$, 95% CI [-0.13, 0.02].

Similarly, participants' interest in joining a dating website with potential partners high in sense of humor did not differ across the strong (*M* = 7.26, *SD* = 1.69) vs. weak (*M* = 7.17, *SD* = 1.76) functional preferences conditions, *t*(2541) = 1.25, *p* = .211, *d* = 0.05, 95% CI [-0.03, 0.13].

Mediational model

We preregistered that we would test whether the ideal partner preference for youthfulness statistically mediated the effect of condition on the partner youthfulness estimate. Using the component approach recommended by Yzerbyt et al. (2018), we found evidence for mediation: The effect of condition on the mediator (ideal for youthfulness) was significant, $\beta = .07$, *t*(2541) = 3.44, *p* < .001, and the effect of the mediator (ideal for youthfulness) on the dependent measure was also significant (partner youthfulness), $\beta = .515$, *t*(2540) = 30.25, *p* < .001. The direct effect of the independent variable (condition) on the dependent measure (partner youthfulness) was not significant when controlling for ideal, $\beta = .004$, *t*(2540) = 0.25, *p* = .805.

Moderation by Satisfaction of the Effectiveness of the Manipulation

Finally, we preregistered that we would test to see whether relationship satisfaction moderated the association between condition and participants' self-reported ideal partner preference for youthfulness (manipulation check). The main effect of condition was significant, $\beta = 0.07$, *t*(2539) = 3.56, *p* < .001, the main effect of satisfaction was significant, $\beta = 0.09$, *t*(2539) = 4.73, *p* < .001, and the interaction was not significant, $\beta = 0.02$, *t*(2539)= 1.04, *p* = .301. In other words, the manipulation boosted the manipulation check approximately the same for all participants, regardless of their level of relationship satisfaction.

Moderated Motivated Perception (Correlational)

We also preregistered that we would test to see whether relationship satisfaction moderated the association between participants' self-reported ideal partner preference for youthfulness (rather than condition, as reported in the main manuscript as RQ3b) and the extent to which they thought their partner was high in youthfulness. In this regression, all three variables were standardized. The main effect of ideals for youthfulness was significant, $\beta = 0.49$, $t(2539) = 29.71$, $p < .001$, the main effect of satisfaction was significant, $\beta = 0.24$, $t(2539) = 13.96$, $p < .001$, and the interaction was significant, $\beta = 0.05$, $t(2539)= 3.23$, $p = .001$. In other words, there was evidence for this "correlational test" of the moderated motivated perception account, but the analysis in the main manuscript did not find causal evidence for the moderated motivated perception account.

Estimating Youthfulness in New Faces

Finally, we preregistered that we would test whether the average estimate of youthfulness in the 12 new face stimuli would differ by condition. We did not find evidence that the estimates were different between the strong ($M = 6.72$, $SD = 1.41$) and weak ($M = 6.70$, $SD = 1.41$) functional preference for youthfulness conditions, $t(2541) = 0.31$, $p = .755$, $d = 0.01$, 95% CI [-0.07, 0.09].

Discussion

In this study, we told participants the real name of the trait instead of a made-up label. For this reason, similarly to Study S1b, it was harder to move participants' ideals about the trait, presumably because they already had stronger preconceived notions of how much they liked it. Thus, the manipulation was weaker (e.g., $d = 0.14$ in this study vs. $d = 0.77$ in Study 2) and we may have been underpowered to detect RQ1, RQ2, RQ3a, RQ3b and RQ5[19]. We did, however, find support for the situation selection account: The manipulation caused participants to want to join a website with younger partners.

[19] We then proceeded to test these effects with a stronger manipulation; the result is Study 3 as reported in the main manuscript.

Task Performance

Even though participants' performance on the DateFest game was not directly relevant to our research questions, we provide the data below for the interested reader. If indeed functional preferences were stronger in the strong functional preferences condition, we should expect participants to have an easier time learning the association between trials and the focal trait, as well as perform better in the game (i.e., earn more points). The mean and standard deviations refer to the number of times participants said 'yes' to a good date (i.e., dates that earned 10 points) or to a bad date (i.e., dates that subtracted 10 points), as specified below. The number of times they said 'yes' is always out of 12 dates, since there were 12 good dates and 12 bad dates in the 24 trials.

Study 1. Participants in the strong functional preferences condition said yes to more good dates (M = 8.76, SD = 2.17 vs. M = 7.89, SD = 1.73) than in the weak functional preferences condition, $t(401) = 4.47$, $p < .001$, d = 0.45, 95% CI [0.25, 0.64]. Again as expected, they also said yes less often to bad dates (M = 2.18, SD = 1.94 vs. M = 3.18, SD = 2.11; $t(401) = -4.95$, $p < .001$, d = -0.49, 95% CI [-0.69, -0.30]) and earned more points in the task (M = 115.18, SD = 30.71 vs. M = 97.08, SD = 24.33; $t(401) = 6.71$, $p < .001$, d = 0.67, 95% CI [0.47, 0.87]).

Study 2. Participants in the strong functional preferences condition said yes to more good dates (M = 9.54, SD = 1.86 vs. M = 7.12, SD = 1.93) than in the weak functional preferences condition, $t(1637) = 25.89$, $p < .001$, d = 1.28, 95% CI [1.17, 1.39]. Again as expected, they also said yes less often to bad dates (M = 1.04, SD = 1.51 vs. M = 3.42, SD = 1.91; $t(1637) = -28.06$, $p < .001$, d = -1.39, 95% CI [-1.49, -1.28]) and earned more points in the task (M = 134.81, SD = 25.06 vs. M = 84.91, SD = 25.17; $t(1637) = 40.21$, $p < .001$, d = 1.99, 95% CI [1.87, 2.11]).

Study 3. Participants in the strong functional preferences condition said yes to more good dates (*M* = 8.92, *SD* = 2.17 vs. *M* = 6.98, *SD* = 1.99) than in the weak functional preferences condition, *t*(2330) = 22.47, *p* < .001, *d* = 0.93, 95% CI [0.85, 1.02]. Again as expected, they also said yes less often to bad dates (*M* = 0.99, *SD* = 1.29 vs. *M* = 2.82, *SD* = 1.86; *t*(2330) = -27.57, *p* < .001, *d* = -1.14, 95% CI [-1.23, -1.05]) and earned more points in the task (*M* = 129.21, *SD* = 24.70 vs. *M* = 90.54, *SD* = 25.06; *t*(2330) = 37.52, *p* < .001, *d* = 1.55, 95% CI [1.46, 1.65]).

Study S1a. This study had two complexity conditions (high and low), and analyses were done separately.

In the high complexity condition, participants in the strong functional preferences condition said yes to more good dates (*M* = 8.50, *SD* = 2.10 vs. *M* = 7.81, *SD* = 1.95) than in the weak functional preferences condition, *t*(426) = 3.47, *p* < .001, *d* = 0.34, 95% CI [0.14, 0.53]. Again as expected, they also said yes less often to bad dates (*M* = 1.90, *SD* = 1.85 vs. *M* = 2.84, *SD* = 1.82; *t*(426) = -5.30, *p* < .001, *d* = -0.51, 95% CI [-0.70, -0.32]) and earned more points in the task (*M* = 115.69, *SD* = 26.98 vs. *M* = 99.76, *SD* = 22.21; *t*(426) = 6.65, *p* < .001, *d* = 0.64, 95% CI [0.45, 0.84]).

In the low complexity condition, participants in the strong functional preferences condition said yes to more good dates (*M* = 7.64, *SD* = 2.23 vs. *M* = 6.34, *SD* = 2.07) than in the weak functional preferences condition, *t*(401) = 6.04, *p* < .001, *d* = 0.60, 95% CI [0.40, 0.80]. Again as expected, they also said yes less often to bad dates (*M* = 2.36, *SD* = 1.62 vs. *M* = 4.51, *SD* = 2.05; *t*(401) = -11.66, *p* < .001, *d* = -1.16, 95% CI [-1.37, -0.95]) and earned more points in the task (*M* = 102.17, *SD* = 24.82 vs. *M* = 68.30, *SD* = 22.97; *t*(401) = 14.21, *p* < .001, *d* = 1.42, 95% CI [1.20, 1.63]).

Study S1b. Participants in the strong functional preferences condition said yes to more good dates (*M* = 8.48, *SD* = 2.08 vs. *M* = 7.59, *SD* = 2.06) than in the weak functional preferences condition, *t*(826) = 6.18, *p* < .001, *d* = 0.43, 95% CI [0.29, 0.57]. Again as expected, they also said yes less often to bad dates (*M* = 2.13, *SD* = 1.88 vs. *M* = 2.71, *SD* = 2.03; *t*(826) = -4.30, *p* < .001, *d* = -0.30, 95% CI [-0.44, -0.16]) and earned more points in the task (*M* = 113.47, *SD* = 27.09 vs. *M* = 98.73, *SD* = 24.33; *t*(826) = 8.22, *p* < .001, *d* = 0.57, 95% CI [0.43, 0.71]).

Study S2. Participants in the strong functional preferences condition said yes to more good dates (*M* = 8.50, *SD* = 2.15 vs. *M* = 7.97, *SD* = 1.95) than in the weak functional preferences condition, *t*(2541) = 6.54, *p* < .001, *d* = 0.26, 95% CI [0.18, 0.34]. Again as expected, they also said yes less often to bad dates (*M* = 2.18, *SD* = 2.08 vs. *M* = 2.99, *SD* = 2.25; *t*(2541) = -9.41, *p* < .001, *d* = -0.37, 95% CI [-0.45, -0.29]) and earned more points in the task (*M* = 112.61, *SD* = 28.28 vs. *M* = 99.77, *SD* = 26.80; *t*(2541) = 11.76, *p* < .001, *d* = 0.47, 95% CI [0.39, 0.55]).

Correlations between Main Variables

Table C1

Study S2 Descriptive Statistics and Correlations

Variable	*M (SD)*	1	2	3	4	5	6	7
1. Ideals for Youthfulness	5.85 (1.77)	-						
2. Estimate of Youth. in partner	8.29 (2.17)	.52***	-					
3. Relationship Satisfaction	7.52 (1.46)	.09***	.27***	-				
4. Interest on website - Youth.	5.17 (2.38)	.66***	.38***	.06**	-			
5. Ideals for sense of humor	8.01 (1.04)	.02	.16***	.28***	-.07***	-		
6. Estimate of soh in partner	9.53 (2.04)	.03	.34***	.43***	-.03	.45***	-	
7. Interest on website - soh	7.22 (1.73)	.05*	.08***	.12***	.29***	.43***	.22***	-

**p* < .05. ***p* < .01., ****p* < .001; *soh* = sense of humor, *youth* = youthfulness.

Table C2

Study 1 Descriptive Statistics and Correlations

Variable	*M (SD)*	1	2	3	4	5	6	7
1. Ideals for Reditry	6.05 (1.98)	-						
2. Estimate of Reditry in partner	8.37 (2.30)	.54***	-					
3. Relationship Satisfaction	7.54 (1.60)	.13**	.42***	-				
4. Interest on website - Reditry	5.78 (2.35)	.61***	.44***	.12*	-			
5. Ideals for soh	8.05 (1.02)	-.04	.11*	.17***	-.07	-		
6. Estimate of soh in partner	9.72 (1.92)	.10*	.42***	.49***	.05	.41***	-	
7. Interest on website - soh	7.24 (1.92)	-.06	.02	.12*	.28***	.42***	.22***	-

$^{*}p$ < .05. $^{**}p$ < .01., $^{***}p$ < .001; *soh* = sense of humor.

Table C3

Study 2 Descriptive Statistics and Correlations

Variable	*M (SD)*	1	2	3	4	5	6	7
1. Ideals for Reditry	5.82 (2.08)	-						
2. Estimate of Reditry in partner	8.06 (2.43)	.53***	-					
3. Relationship Satisfaction	7.53 (1.64)	.04	.32***	-				
4. Interest on website - Reditry	5.31 (2.54)	.51***	.41***	.06*	-			
5. Ideals for soh	7.90 (1.17)	-.07**	.02	.12***	-.07**	-		
6. Estimate of soh in partner	9.71 (2.11)	.00	.30***	.47***	-.01	.40***	-	
7. Interest on website - soh	7.12 (1.99)	-.05*	.00	.10***	.28***	.39***	.24***	-

$^{*}p$ < .05. $^{**}p$ < .01., $^{***}p$ < .001; *soh* = sense of humor.

Table C4

Study 3 Descriptive Statistics and Correlations

Variable	*M (SD)*	1	2	3	4	5	6	7
1. Ideals for Youthfulness	5.58 (1.61)	-						
2. Estimate of Youth. in partner	8.35 (2.16)	.52***	-					
3. Relationship Satisfaction	7.72 (1.51)	.07**	.26***	-				
4. Interest on website – Youth.	4.36 (2.29)	.54***	.34***	.05**	-			
5. Ideals for soh	8.11 (0.96)	.05*	.11***	.13***	.00	-		
6. Estimate of soh in partner	10.05 (1.86)	.10***	.38***	.46***	.04*	.40***	-	
7. Interest on website - soh	6.93 (2.04)	.03	.04	.09***	.38***	.36***	.16***	-

**p* < .05. ***p* < .01., ****p* < .001; *soh* = sense of humor.

Stimuli Used in the DateFest Tasks

Table C5

Stimuli used in Study 1

CFD id	Trial	Reditry	Sense of Humor	Condition
WM-255	dislike	27	53	weak
WM-018	dislike	34	61	weak
WM-016	dislike	36	44	weak
WM-225	dislike	37	21	weak
WM-211	dislike	43	85	weak
WM-238	dislike	47	49	weak
WM-247	dislike	56	59	weak
WM-232	dislike	64	69	weak
WM-006	dislike	76	37	weak
WM-039	dislike	84	45	weak
WM-001	dislike	94	62	weak
WM-024	dislike	94	50	weak
WM-227	like	41	83	weak
WM-251	like	44	97	weak
WM-205	like	52	78	weak
WM-017	like	63	53	weak

WM-213	like	74	117	weak
WM-202	like	80	82	weak
WM-200	like	91	90	weak
WM-214	like	94	107	weak
WM-236	like	101	66	weak
WM-037	like	102	76	weak
WM-231	like	109	91	weak
WM-239	like	116	84	weak
WM-255	dislike	27	53	strong
WM-018	dislike	34	61	strong
WM-016	dislike	36	44	strong
WM-225	dislike	37	21	strong
WM-227	dislike	41	85	strong
WM-211	dislike	43	49	strong
WM-251	dislike	44	59	strong
WM-238	dislike	47	69	strong
WM-205	dislike	52	37	strong
WM-017	dislike	63	45	strong
WM-213	dislike	74	62	strong
WM-202	dislike	80	50	strong
WM-247	like	56	83	strong
WM-232	like	64	97	strong
WM-006	like	76	78	strong
WM-039	like	84	53	strong
WM-200	like	91	117	strong
WM-001	like	94	82	strong
WM-214	like	94	90	strong
WM-024	like	94	107	strong
WM-236	like	101	66	strong
WM-037	like	102	76	strong
WM-231	like	109	91	strong
WM-239	like	116	84	strong

Table C6

Stimuli used in Studies 2 and 3

CFD id	Trial	Reditry	Sense of Humor	Condition
WM-223	dislike	12	63	weak
WM-018	dislike	34	59	weak
WM-249	dislike	26	54	weak
WM-236	dislike	101	24	weak
WM-211	dislike	43	47	weak
WM-230	dislike	106	84	weak
WM-234	dislike	65	57	weak
WM-208	dislike	98	67	weak
WM-253	dislike	77	46	weak
WM-202	dislike	80	42	weak
WM-214	dislike	94	61	weak
WM-235	dislike	56	48	weak
WM-204	like	38	94	weak
WM-205	like	52	101	weak
WM-006	like	76	82	weak
WM-238	like	47	17	weak
WM-256	like	61	120	weak
WM-206	like	87	80	weak
WM-217	like	115	47	weak
WM-250	like	98	111	weak
WM-225	like	37	90	weak
WM-227	like	41	86	weak
WM-231	like	109	95	weak
WM-239	like	116	98	weak
WM-223	dislike	12	63	strong
WM-018	dislike	34	59	strong
WM-249	dislike	26	42	strong
WM-225	dislike	37	24	strong
WM-227	dislike	41	84	strong
WM-211	dislike	43	47	strong
WM-256	dislike	61	57	strong
WM-238	dislike	47	67	strong
WM-205	dislike	52	46	strong
WM-234	dislike	65	54	strong
WM-204	dislike	38	61	strong
WM-235	dislike	56	48	strong
WM-214	like	94	47	strong

WM-253	like	77	101	strong
WM-006	like	76	82	strong
WM-208	like	98	17	strong
WM-202	like	80	120	strong
WM-206	like	87	86	strong
WM-217	like	115	94	strong
WM-250	like	98	111	strong
WM-236	like	101	90	strong
WM-230	like	106	80	strong
WM-231	like	109	95	strong
WM-239	like	116	98	strong