RUNNING HEAD: ATTRACTIVENESS AND DESIRE

Sex-Differentiated Effects of Physical Attractiveness on Romantic Desire: A Highly Powered, Preregistered Study in a Photograph Evaluation Context

Paul W. Eastwick and Leigh K. Smith

University of California, Davis

In Press, Comprehensive Results in Social Psychology

Corresponding author:

Paul W. Eastwick
Department of Psychology
University of California, Davis
1 Shields Ave
Davis, CA 95616
email: eastwick@ucdavis.edu
phone: 773 484 3878

Abstract

A large, controversial literature has examined the hypothesis that the attractiveness of potential partners predicts romantic desire more strongly for men than for women. Nevertheless, prior studies demonstrating this sex difference in photograph-evaluation contexts have used extremely small samples of stimuli, which is as detrimental to statistical power as the use of small samples of participants. The current registered report used very large samples of both participants (N=1,204) and stimuli photographs (N=593) to test the sex difference in the attractiveness-desire association. The sex difference emerged with objective assessments of attractiveness from independent raters (approximately q = .13, a small effect) but not with participants' own assessments of attractiveness (q = .00). Various other moderators that have been summoned to explain cross-study variability in prior research received no support (e.g., the sex difference was not larger in serious relationship contexts, the low-to-moderate range of attractiveness, etc.). Surprisingly, in the small sample of participants who were attracted to same-sex individuals, the attractiveness-desire association was stronger for women than men—the opposite of the sex difference anticipated by prior mate preferences research. This study provides effect-size benchmarks for studies of sex differences and highlights the importance of stimulus sampling when documenting replicable effects.

Keywords: Attractiveness; sex differences; romantic attraction; evolutionary psychology; stimulus sampling

Main text word count: 10,086

Sex-Differentiated Effects of Physical Attractiveness on Romantic Desire:

A Highly Powered, Preregistered Study in a Photograph Evaluation Context

Men tend to indicate that they value physical attractiveness in a mate more highly than women when imagining their ideal romantic partner; for many decades, researchers have studied this sex difference extensively across the fields of close relationships (Sprecher, Sullivan, & Hatfield, 1994), family studies (Hill, 1945), and evolutionary psychology (Buss, 1989). This sex difference is evident when men and women consider all types of ideal partners—from marriage partners to dating partners to sex partners (Buunk, Dijkstra, Fetchenhauer, & Kenrick, 2002; Kenrick, Sadalla, Groth, & Trost, 1990; Li & Kenrick, 2006; Regan, Levin, Sprecher, Christopher, & Cate, 2000). Furthermore, it emerges reliably in meta-analyses (Feingold, 1990) and representative samples collected in complex modern societies (Sprecher et al., 1994; cf. W. Wood & Eagly, 2002).

Partner preferences are meaningful to the extent that they have consequences for people's evaluations of (and subsequent behavior toward) actual mating partners in the real world (Simpson, Fletcher, & Campbell, 2001). Thus, the most straightforward way that the sex difference in the preference for attractiveness could exhibit such predictive validity is as follows: The attractiveness of a partner should predict people's romantic evaluations of that partner more strongly for men than for women (Eastwick & Finkel, 2008; Eastwick, Luchies, Finkel, & Hunt, 2014a). This postulate—and the evidence that bears on it—has proven controversial in recent years. A large meta-analysis (N = 29,414; Eastwick et al., 2014a) has suggested that there is no sex difference in the association of physical attractiveness with romantic evaluations in either initial attraction contexts (e.g., attractiveness predicts desire for confederates or speed-daters at $r \approx .50$ for men and women) or established close relationship contexts (e.g., attractiveness predicts

relationship satisfaction at $r \approx .35$ for men and women). Yet in the wake of this meta-analysis, some scholars have emphasized data that supports the existence of these sex differences (e.g., Li et al., 2013; Meltzer, McNulty, Jackson, & Karney, 2014a; cf. Eastwick, Neff, Finkel, Luchies, & Hunt, 2014), whereas other scholars have subsequently argued that evolutionary theories do not predict that these sex differences should emerge in the first place (Schmitt, 2014; cf. Eastwick, Luchies, Finkel, & Hunt, 2014b).

Despite these empirical and theoretical disputes, both the skeptics and the skeptics-of-the-skeptics agree on a key empirical tenet: Attractiveness does (and should) predict men's evaluations more strongly than women's evaluations in contexts where people evaluate potential partners whom they have never met (Eastwick et al., 2014a). That is, setting aside the controversy surrounding the evidence for the replicability of the physical attractiveness sex difference in initial attraction and close relationships contexts, the sex difference seems to be quite robust when people evaluate photographs and similar hypothetical partners (e.g., descriptions of potential partners, personal ads). Indeed, in the published literature, there are numerous successful demonstrations of this sex difference (e.g., de Vries, 2010; Li et al., 2013, Study 2; Townsend & Levy, 1990b; Townsend & Roberts, 1993; Wenzel & Emerson, 2009) as well as an early meta-analysis of photograph evaluations suggesting that the attractiveness-evaluation effect size for men was approximately double the effect size for women (Part 5 of Feingold, 1990). In light of this evidence, many theories of ideal partner preferences have needed to incorporate the proposition that partner preferences predict evaluations better in hypothetical contexts than contexts in which people are evaluating attraction to partners they have met face-to-face (Eastwick, Finkel, & Eagly, 2011).

Before we continue to build frameworks for human mating that assume that the sex difference in the effect of attractiveness on romantic desire in hypothetical settings is robust, it is worth pausing to consider the strength of this evidence. Scholars have recently begun to appreciate the extent to which literatures with low statistical power generally yield less reliable evidence for or against hypotheses (Button et al., 2013; Ioannidis, 2005; Lakens & Evers, 2014), and as it turns out, many of the studies demonstrating the attractiveness sex difference in hypothetical settings do not conform to contemporary standards of statistical power. Intriguingly, this issue has likely gone unnoticed because it is not the number of *participants* in these studies that is concerning but rather the extremely small number of *stimuli* in these studies (i.e., the attractive and unattractive photographs that participants evaluate). The routine use of small samples of stimuli can have unfortunate consequences for replicability—just like the routine use of small samples of participants—because parameter estimates relying on a small number of stimuli are unstable and are likely to fail to generalize to other populations of stimuli (Judd, Westfall, & Kenny, 2012, 2017; Wells & Windschitl, 1999; Westfall, Judd, & Kenny, 2015; Westfall, Kenny, & Judd, 2014). To address this issue, the current study examines the sex difference in the predictive effects of physical attractiveness using a large sample of participants as well as a large sample of photographic stimuli drawn from the publically available Chicago face database (Ma, Correll, & Wittenbrink, 2015).

Existing Studies of Physical Attractiveness When Participants Evaluate Photographs

Numerous studies require participants to evaluate photographs which have been selected by the researchers to represent different levels of attractiveness. Furthermore, many of these studies state the hypothesis (and search for evidence) that the attractiveness of the photographs should affect men's romantic evaluations of the photographs more strongly than it affects

women's evaluations. For example, one early demonstration of this effect asked participants to rate three opposite-sex photographs that were judged by an independent set of raters to be high, medium, or low on physical attractiveness (Townsend & Levy, 1990b). The partners' attractiveness positively predicted participants' evaluations on several items (e.g., "I would like to go on a date with a person like this," "I would be willing to have a serious relationship with a person like this that could lead to marriage"), and this effect was stronger for men's than women's evaluations.

A closer inspection of this literature reveals that many studies documenting this sex difference use only a small number of stimuli—that is, participants rate a small number of opposite-sex photographs as potential partners, and those same photographs are used across all participants. For example, Townsend and Levy (1990b) find the sex difference with $n = 3$ photographs of each sex (i.e., one-low, one-medium, and one high-attractiveness photograph), de Vries (2010) finds the sex difference with $n = 4$ photographs of each sex (i.e., two low- and two high-attractiveness photographs), and Li et al. (2013, Study 2) finds the sex difference with $n = 6$ photographs of each sex (i.e., two low-, two medium-, and two high-attractiveness photographs). Presumably, one goal of these studies is to generalize from a sample of potential partners (i.e. the photographs used in the studies) to a broader population of potential partners (i.e., the potential partners that participants might encounter in their lives). Given this goal, the photographs in these studies should be considered a random factor, and a sample size of $n = 6$ for a random stimulus factor is as problematic as running a study on $N = 6$ participants (Judd et al., 2012, 2017; Wells & Windschitl, 1999; Westfall et al. 2015; Westfall et al., 2014). Thus, despite the fact that these studies laudably used hundreds of participants, statistical power was nevertheless strikingly low.

There are other studies in the published literature in which participants rate photographs that vary in attractiveness, yet the sex difference was not a focal hypothesis. These studies often use larger samples of stimuli and report (usually in a brief aside or footnote) that participant sex did not moderate the effect of attractiveness on stimuli ratings. For example, Lewandowski, Aron, and Gee (2007) did not find the sex difference (G. Lewandowski, personal communication, January 26, 2017) with n = 36 photographs of each sex (i.e., 12 low-, 12 medium-, and 12 high-attractiveness photographs), Montoya (2008) did not find the sex difference with n = 14 photographs of each sex (i.e., 2 photographs at each of 7 levels of attractiveness), and Ritter, Karremans, and van Schie (2010) did not find the sex difference with n = 80 photographs of each sex (i.e., 40 low- and 40 high-attractiveness photographs). In other words, when researchers use large samples of stimuli and are not explicitly looking for the sex difference, the size of the sex difference tends to be small and indistinguishable from zero (see also Olderbak, Malter, Wolf, Jones, & Figueredo, 2017).

There are three studies that do not conform to the general trend whereby the physical attractiveness sex difference emerges in small but not large samples of stimuli: two studies of online dating (Hitsch, Hortacsu, & Ariely, 2010; Lee, Loewenstein, Ariely, Hong, & Young, 2008) and one study of a traditional matchmaking service (de Vries, Swenson, & Walsh, 2007). These datasets were considerably larger, consisting of hundreds (de Vries et al., 2007) or thousands (Hitsch et al., 2010; Lee et al., 2008) of participants and stimuli. All three found that dating site users received more meeting requests to the extent that they were attractive, and this effect was stronger for women's photos (i.e., men messaging women) than for men's photos.

Nevertheless, these purported demonstrations of the physical attractiveness sex difference are not especially definitive. These studies all examined a naturalistic context in which users

could decide which photos of themselves to share with dating site users, and this element of the procedure opens the opportunity for several possible confounds to emerge. For example, imagine that (a) making a "flirty face" generally elicits more messages on online dating sites (Rudder, 2010), and (b) the association between one's own attractiveness and posting a "flirty face" photo is stronger for women than for men (i.e., attractive women but not attractive men are especially likely to make a "flirty face"). In this plausible scenario, attractiveness would elicit more messages from men than from women not because attractiveness per se is more desirable to men but because there are sex differences in the desirable, message-eliciting qualities (e.g., "flirty-face-ness") that accompany attractive photographs. A clearer test of the attractiveness sex difference that avoids such confounds would entail the use of standardized stimuli that depict real people but do not allow the stimuli themselves to choose how they want to appear in the photograph.

In summary, there is ambiguous evidence for the sex difference in the predictive effects of attractiveness in photograph-evaluation contexts. When considering the published literature on this effect as a whole, replicability is perhaps the most pressing scientific desideratum in need of reinforcement (Finkel, Eastwick, & Reis, in press). But in this case, such a reinforcement will not be persuasive if it comes at the cost of internal validity (e.g., if stimuli can select their own photographs).

The Benefits of Sampling Large Numbers of Stimuli

Researchers are well acquainted with the idea that participants are a random factor in experimental designs: Typically, we wish to generalize from the sample of participants to a population of possible participants, and large samples provide greater confidence that the effect size in a sample is comparable to the effect size in a population. New statistical approaches have

highlighted how the stimuli that participants evaluate as part of an experimental design are also random factors (Judd et al., 2012, 2017; Wells & Windschitl, 1999; Westfall et al. 2015, Westfall et al. 2014). That is, there are many cases where researchers also wish to generalize from a sample of stimuli to a larger population of stimuli, and thus, it is equally important that researchers employ large samples of stimuli. Otherwise, a finding might be true yet systematically restricted to the small number of stimuli that the researchers happened to select, and thus any significant or nonsignificant effect that emerges will fail to generalize to stimuli outside the chosen set in a meaningful way.

This sampling issue is particularly acute when researchers investigate opposite-sex attraction, because men and women typically evaluate different stimuli in these designs. Thus, to the extent that researchers design stimuli that vary on a particular independent variable of interest, it is very challenging to determine that the independent variable is both equally strong and equally unconfounded for men and for women. Consider the selection of two low attractiveness (e.g., Amanda and James) and two high attractiveness (e.g., Rachel and Brian) individuals of each sex from a larger pool of possible targets. In the low-target *n* studies described above, researchers would ensure that Amanda and James possess equivalent attractiveness (e.g., 2.4 on a 7-point scale) and that Rachel and Brian possess equivalent attractiveness (e.g., a 5.7 on a 7-point scale; see Li et al., 2013, Study 2 for similar values).

Although this design does manipulate attractiveness, it has two major limitations. The first limitation is that these attractive-unattractive pairs could conceivably differ on any other quality that participants can perceive from a face (e.g., unhappy, sociable, emotionally stable, mean, boring, aggressive, weird, intelligent, confident, caring, egotistic, responsible, trustworthy, dominant; Oosterhof & Todorov, 2008). It would be very easy for researchers to have

unintentionally selected photographs in which the two male faces differed on several of these desirable qualities (e.g., Brian is much more trustworthy than James) more than the two female faces differed (e.g., Rachel and Amanda are equally trustworthy), thus creating a stronger "attractiveness" manipulation for one sex. With so few stimuli, confounds could be very difficult to rule out. The second limitation is that, with respect to attractiveness specifically, the validity of this procedure requires that the male and female distributions of attractiveness align such that a 2.4 for a man is roughly equivalently unattractive as a 2.4 for woman. This assumption is false: Women are more attractive than men on average (Fletcher, Kerr, Li, & Valentine, 2014; Rudder, 2014; D. Wood & Brumbaugh, 2009), and therefore targets that are matched on numerical values occupy different areas of the attractiveness distribution across the two sexes (Eastwick et al., 2014b). In the Chicago face database (Ma et al., 2015), a 2.4 (on a 7-point attractiveness scale) for a man would be approximately the 20th percentile (i.e., James), whereas a 2.4 for a woman is approximately the 10th percentile (i.e., Amanda). Given that a 5.7 is above the 99th percentile for both men and women (i.e., Rachel and Brian), the Amanda-James vs. Rachel-Brian attractiveness manipulation is stronger for the female pair than the male pair. These problems can largely be mitigated by sampling large numbers of stimuli that capture the full range of attractiveness in the population to which the researcher wishes to generalize.

The Current Research

The current study tested whether the association of attractiveness with romantic desire is sex-differentiated in a traditional hypothetical setting by asking male and female participants to evaluate several hundred photographs from the Chicago face database (Ma et al., 2015). The current version of the database contains photographs of approximately 300 men and 300 women, all of whom posed with a neutral expression, looked directly at the camera, and wore a plain grey

t-shirt. Also, all the people depicted in the Chicago face database are everyday people (i.e., not models) and therefore represent a reasonable range of potential partners that participants might be likely to meet in real life.

Participants in the current study reported their romantic desire for each of the opposite-sex photographs, and they also provided a personal judgment of each target's attractiveness. The Ma et al. (2015) manuscript also contains data on independent raters' evaluations of the attractiveness of the photographs (i.e., the "norming data"), and not surprisingly, the women were rated approximately .5 *SD*s higher than the men on average (Figure 1).

This study tested the following primary hypothesis: In a photograph-rating context, the association of attractiveness with romantic desire is stronger for men than for women. Photograph attractiveness was assessed in two ways: the participant's personal rating of each photograph and the norming-data attractiveness rating of each photograph. We also tested this sex difference in several theoretically meaningful subsidiary analyses that may be especially likely to reveal a stronger association for men than for women (e.g., tests using a "serious relationship" dependent measure alone; tests on targets in the bottom half of the attractiveness distribution; see Analysis Plan section below). For each analysis, we conducted both significance tests (i.e., to attempt to reject the null hypothesis of no sex difference) as well as equivalence tests (i.e., to attempt to reject the null hypothesis of a meaningful sex difference; Lakens, in press).

Method

Participants

Participants received $2.00 for completing a ~40 min photograph rating task on Mechanical Turk. Participants were excluded from all subsequent analyses if (a) they did not

finish the survey (i.e., the "Finished" column in the Qualtrics data file contains a "0"; n = 340), (b) they gave an identical numerical response to all photographs on any of the three items (because the participant then provides no within-person association between attractiveness and desire; n = 35), (c) they provided a response other than male or female to the question asking about their sex (n = 4), or (d) they failed either of the two attention checks interspersed throughout the rating task (n = 420). Participants who responded that they are primarily attracted to same-sex partners rated same-sex faces, but their data are included in the Auxiliary Analyses section only.

We originally planned to recruit a sample of 700 male and 700 female participants from Mechanical Turk with the goal of obtaining a usable sample of 600 men and 600 women who were primarily attracted to members of the opposite-sex. (For a justification of these estimates, see Power Analysis section below.) We paused data collection after recruiting 70 men and 70 women (10% of the anticipated final sample) to ensure the survey was running smoothly. At this point, we determined that we would need to exclude a larger percentage of participants than we anticipated (given the above exclusion criteria), and the editor approved raising the recruitment target to 800 men and 800 women. We paused data collection a second time after recruiting 1593 participants to determine how many usable participants remained after applying our exclusion criteria. At this point, we needed an additional ~30 usable men and ~50 usable women attracted to opposite-sex partners to reach our target of 600 men and 600 women, and so the editor approved the collection of an additional 65 men and 85 women. We ended data collection after recruiting and paying 1747 participants through Mechanical Turk.

This total left us with a usable sample of n = 609 men and n = 595 women attracted to opposite-sex partners and n = 46 men and n = 57 women attracted to same-sex partners who

completed the study and passed all data quality checks. (All of the "identical numerical response" excluded participants and most of the "failed attention check" excluded participants also counted toward the paid participant total.) Our sample of participants attracted to same-sex individuals is small, which is a consequence of the fact that (a) the study was powered to detect effects among participants attracted to opposite-sex partners and (b) individuals who identify as nonheterosexual comprise a modest percentage of the Mechanical Turk population (i.e., ~10%; Coffman, Coffman, & Ericson, 2016). We did not substantively examine the data until this total sample had been collected.

Materials and Procedure

After learning about the study on the Mechanical Turk website, participants first completed a screener questionnaire asking their sex (with the response options "male," "female," and "other" with a textbox) and age (opposite-sex raters M_{age} = 32.5, SD = 6.1; same-sex raters M_{age} = 30.9, SD = 6.1). Given that 95% of the photographs in the Chicago face database are aged 41 and younger, participants were screened out of the study if they were 46 years old or older. Also, participants were screened out of the study if the survey had reached the quota for their selected sex (total screened out n = 546).

Eligible participants read the consent form and responded to an item asking whether they are primarily romantically attracted to members of the same sex or the opposite sex. Participants selecting "opposite sex" were shown all of the opposite-sex photographs in the Chicago face database, whereas participants selecting "same sex" were shown all of the same-sex photographs in the database. Participants also completed an item assessing their race/ethnicity with the following response options: "African-American, Black, African, Caribbean" (opposite-sex raters 9.9%, same-sex raters 11.7%), "Asian-American, Asian, Pacific Islander" (opposite-sex raters

6.8%, same-sex raters 4.9%), "European-American, Anglo, Caucasian" (opposite-sex raters 71.2%, same-sex raters 72.8%), "Hispanic-American, Latino(a), Chicano(a)" (opposite-sex raters 7.1%, same-sex raters 6.8%), "Native-American, American Indian" (opposite-sex raters 1.0%, same-sex raters 0.0%), "Bi-racial, Multi-racial" (opposite-sex raters 3.6%, same-sex raters 2.9%) and "Other (Please indicate in text box)" (opposite-sex raters 0.4%, same-sex raters 1.0%). They also indicated whether they are currently involved in a romantic relationship with the (mutually exclusive) response options "Married" (opposite-sex raters 43.1%, same-sex raters 22.5%), "Unmarried, but in a serious relationship" (opposite-sex raters 20.8%, same-sex raters 30.4%), and "Single" (opposite-sex raters 36.0%, same-sex raters 47.1%). Finally, they completed a brief set of measures that included the following three items: "To what extent do you desire 'physical attractiveness' in an ideal romantic partner?", "To what extent does a person's 'physical attractiveness' increase the likelihood that you will want to go on a date with them?", and "To what extent does a person's 'physical attractiveness' increase the likelihood that you will want to have a serious relationship with them that could lead to marriage?" Participants completed these *Stated Preference* (opposite-sex raters $\alpha = .88$; same-sex raters $\alpha = .93$) items on a scale from 1 (*not at all*) to 9 (*a great deal*). Consistent with previous findings (e.g., Bailey, Gaulin, Agyei, & Gladue, 1994; Buss, 1989), straight men rated this construct more positively than straight women, $M_{Men} = 7.33$, $SD_{Men} = 1.29$, $M_{Women} = 6.84$, $SD_{Women} = 1.32$, $M_{Difference} = 0.49$, 95% CI (.34, .64), $t(1202) = 6.50$, $p < .001$, $d = .37$, and gay men rated this construct more positively than lesbian women, $M_{Men} = 7.13$, $SD_{Men} = 1.35$, $M_{Women} = 6.36$, $SD_{Women} = 1.61$, $M_{Difference} = 0.77$, 95% CI (.18, 1.36), $t(101) = 2.60$, $p = .011$, $d = .52$.

Participants then learned that their task was to rate ~300 faces on three items: "I find this person extremely physically attractive," "I would be interested in going on a date with this

person" (taken from Li et al., 2013), and "I would be interested in having a serious relationship with this person that could lead to marriage" (taken from Townsend & Levy, 1990b). Participants completed these items on a scale from 1 (*Strongly disagree)* to 7 (*Strongly agree*) with a middle anchor at 4 (*neutral*), and they were told to respond to these items as if they were currently single and interested in a new romantic relationship. The first item served as the participant-rating of *physical attractiveness* and the latter two items served as the *romantic desire* dependent measure (opposite-sex raters $\alpha = .96$, same-sex raters $\alpha = .95$). (A second objective measure of physical attractiveness had already been collected as part of the norming data; Ma et al., 2015, Figure 1).

Women attracted to men rated the 288 male photographs in the Chicago face database that are under age 50, and men attracted to women rated the 305 female photographs in the database that are under age 50. Photographs were presented randomly (specifically, nine or ten photos were included on each page, and for each participant, the photos were randomly ordered on each page and the pages themselves were ordered randomly). After completing the rating task, participants were thanked and given a Mechanical Turk completion code.

In our prior experience, MTurk participants are typically able to complete a photograph rating task of this magnitude in a single sitting (provided that the photographs load quickly and there are no connectivity issues; see also Brown-Iannuzzi, Dotsch, Cooley, & Payne, in press; Hauser & Schwarz, 2015; Klein et al., 2014). Nevertheless, we safeguarded against participant fatigue in two ways. First, two attention check items were interspersed throughout the rating task: An item instructing participants to select the middle response on a series of seven radio buttons arranged horizontally, and an item instructing participants to select "Other" in response to the question "On what continent do you live?" Second, we calculated the extent to which

participants agreed on average in their attractiveness ratings (i.e., the average interrater correlation) for the first 50 photographs (when participants were likely alert and engaged) and the last 50 photographs (when participants might be fatigued) that they encountered. Our analysis plan was as follows: If the difference between these two values is greater than $r = .10$, we would recalculate all analyses on only the first 150 targets that each participant rated, and these analyses would be included as a supplement. The values were: interrater $r = .58$ for the first 50 photographs ($r = .64$ for male raters, $r = .52$ for female raters) and interrater $r = .55$ for the last 50 photographs ($r = .60$ for male raters and $r = .50$ for female raters). This analysis showed little evidence of fatigue (r difference = .03); following our analysis plan, we did not recalculate the findings for the first 150 targets.

Power Analyses

We determined that the small effect size $q = .10$ (i.e., a difference between two correlations or two beta weights of approximately .10; e.g., male $\beta = .25$ and female $\beta = .15$) would be the smallest meaningful sex difference that scholars might wish to detect. According to Cohen (1988), a $q = .10$ is considered "small" and approximately equivalent in magnitude to $d = .20$ and $r = .10$. Values smaller than $q = .10$ are generally of little practical consequence and cannot be detected without extraordinarily large samples (i.e., several thousand participants). Also, $q = .10$ is smaller than the significant sex differences detected than the small-stimulus n studies described above (de Vries, 2010; Li et al., 2013; Townsend & Levy, 1990b).

In the regression analyses below, we used $\beta_{dif} = .10$ (approximately equivalent to $q = .10$) as a benchmark for calculating the power to detect a significant ($p < .05$) difference between the male and female βs for attractiveness. Consistent with a frequentist (i.e., null-hypothesis testing) framework, we conclude that there is a nonzero difference between men and women when the

95% confidence interval for β_{dif} does not include zero. We also used β_{dif} = .10 to set the equivalence region inside which we would conclude that there is no substantive difference between men and women (Lakens, in press); also consistent with a frequentist framework, we conclude that there is no meaningful difference between men and women when the 90% confidence interval for β_{dif} is entirely contained in the interval between -.10 and .10. Note that it is possible in principle for a given analysis to be both statistically significant (i.e., 95% confidence interval for $\beta_{dif} > 0$) and equivalent (i.e., 90% confidence interval for $\beta_{dif} < |.10|$; Lakens, 2017).

The power analysis for this study was graciously conducted by Dr. Jacob Westfall, who served as a reviewer on this manuscript. The details of the power analysis simulations are provided at this link: https://goo.gl/pUJqfV. In simulated datasets where the attractiveness β_{dif} = .10, N = 550 participants were required to achieve 80% power to reject the null hypothesis of no sex difference, and N = 800 participants were required to achieve 90% power. In simulated datasets where the attractiveness β_{dif} = .00, N = 700 participants were required to achieve 80% power to correctly conclude that there is no difference using an equivalence test ($|\beta_{dif}| < .10$), and N = 1200 participants were required to achieve 90% power. Thus, by planning to recruit N = 1,400 (700 men and 700 women), we estimated that we would meet or exceed the 90% power estimate for both the significance and equivalence tests (i.e., 800 and 1,200, respectively) after making the exclusions described in the Participants section above.

Analysis Plan

Data preparation. The dataset was organized such that each row contains a participant's ratings of a single stimulus, with separate columns dedicated to the physical attractiveness rating,

the desire to date rating, and the desire for a serious relationship rating. Analyses were conducted using SAS Proc Mixed using code with the following general structure:

```
proc mixed data = CF.ChicagoFaces method = reml covtest;
class participant_id face_id;
model RomanticDesire = Attractiveness Sex Attractiveness*Sex / solution;
random Intercept Attractiveness /sub= participant_id type = vc;
random Intercept Attractiveness /sub=face_id type = vc;
run;
```
[1]

The two random statements account for the nesting due to repeated measurements at the level of participant (i.e., each participant provides ~300 ratings) and face stimulus (i.e., each face is rated ~700 times), respectively. Both random statements model the slope of attractiveness on romantic desire as well; we planned to remove attractiveness from the stimulus (i.e., face_id) random statement in cases where the models do not converge. (All models converged successfully, so we did not have to remove attractiveness from the random statement.) All continuous variables were standardized separately for each sex for each analysis in order to obtain the standardized β for men and women separately, which is an effect size with a similar interpretation as *r*. Participant sex was coded -.5 = male and .5 = female; thus, the interaction term represents the difference between the male and female β and has a similar interpretation as *q*. If β_{dif} is negative, it means that the association between attractiveness and romantic desire is stronger for men (i.e., the predicted direction of the sex difference).

In all analyses in this article, attractiveness serves as an independent variable. Thus, this research does not speak to the large literature investigating which facial features affect whether or not people evaluate someone as attractive or unattractive in the first place (Cunningham,

[1] In the preregistered analysis plan, both covariance structures were set to type = un. When we ran those models on a Windows PC with 16 Gigs of RAM, we received the error "The SAS System stopped processing this step because of insufficient memory." Thus, we followed the SAS recommendations for relieving processing demands (Kiernan, Tao, & Gibbs, 2012) by changing "un" to "vc." In principle, this change in covariance structure should not affect fixed effect estimates.

1986; Cunningham, Barbee, & Pike, 1990; Little, Jones, & DeBruine, 2011; Perrett et al., 1998). This is a separate literature that, not surprisingly, reveals evidence of many sex differences; in the Chicago Face database norming data, for example, a composite of the masculinity and femininity (reverse-scored) ratings predicts attractiveness for men and women quite differently (i.e., male $r = .21$, $p < .001$; female $r = -.85$, $p < .001$). The present article examines the downstream question of whether the attractiveness construct itself has sex-differentiated consequences for dependent measures such as romantic desire; additional data on the predictors of attractiveness can be found in the original Chicago face database article (Ma et al., 2015).

Results

Pre-registered Analyses

Primary analyses. The primary analyses are presented in Table 1. The first row examines the association of the participant's own rating of the attractiveness of each photo (i.e., *participant report* of attractiveness) with the participant's romantic desire rating for the photo. The fourth row examines the association of the norming data attractiveness rating assigned to each photo (i.e., *objective* attractiveness) with the participant's romantic desire rating for the photo. Theoretically, it is unclear which measure of physical attractiveness should be more likely to reveal a sex difference: a measure that has been filtered through each participant's own subjective construal (i.e., the first row) or a measure that captures the consensus among a set of independent raters (i.e., the fourth row; Eastwick, Neff, et al., 2014; cf. Li & Meltzer, 2015). In this study, the sex difference using the participant report of attractiveness was extremely small and not significant ($\beta_{dif} = .00$), whereas the sex difference using objective attractiveness ratings was significant and larger than the region of equivalence ($\beta_{dif} = -.13$). That is, we can conclude that there is no sex difference in the association of participants' own ratings of attractiveness

with romantic desire, but the association of independent raters' judgments of attractiveness with romantic desire is larger for men than for women.

The remaining rows in Table 1 present these two analyses separately for the date item alone (second and fifth row) and the serious relationship item alone (third and sixth row). Generally speaking, the sex difference in the preference for attractiveness is similarly sized whether people consider an ideal short-term or long-term mating partner (Buss & Schmitt, 1993; Buunk et al., 2002; Eastwick et al., 2014b; Kenrick et al., 1990; Li & Kenrick, 2006; Regan et al., 2000). Nevertheless, some perspectives argue that sex-differentiated effects of physical attractiveness on romantic evaluations will be especially pronounced in the context of long-term, serious relationships (e.g., Kenrick, Trost, Groth, & Sadalla, 1993; Li & Meltzer, 2015; Meltzer et al., 2014a, 2014b; Schmitt, 2014); these perspectives predict that the sex difference might be especially likely to emerge on the serious-relationship item. To address this possibility, we conducted analyses on these two items separately. Analyses on these items did not differ substantively from the primary romantic desire analyses: The sex difference was small and nonsignificant using the participant-report attractiveness ratings for both the "date" and "serious relationship" item dependent measures, whereas the sex difference was significant using the objective attractiveness ratings for both the "date" and "serious relationship" items (β_{dif} = -.13 and β_{dif} = -.12, respectively). In other words, we obtained no evidence to support perspectives that argue that sex differences should be more pronounced in the context of long-term, serious relationships.

Auxiliary analyses. Several auxiliary analyses are presented in Table 2. These auxiliary analyses examined the association of physical attractiveness with romantic desire for theoretically meaningful subsamples of participants. All of the these analyses were conducted

using both the participant's rating of the attractiveness of each photo (top half of the table) and the norming data attractiveness rating of each photo (bottom half of the table). In general, the participant-report analyses (i.e., top half of the table) revealed no sex differences across the subsamples, whereas the objective analyses (with one exception) revealed evidence that the attractiveness association is stronger for men than for women across the subsamples.

The *Own Race* analysis applies to participants who selected "African-American, Black, African, Caribbean," "Asian-American, Asian, Pacific Islander," "European-American, Anglo, Caucasian," and "Hispanic-American, Latino(a), Chicano(a)" for the race question at the beginning of the study. The Chicago face database has separate categories for Asian, Black, Latino(a), and White faces, and so this analysis examined if the sex difference in the preference for physical attractiveness applies specifically to cases where participants rate members of their own race. These analyses did not substantively differ from the primary romantic desire analyses (i.e., effect size near zero for the participant-report attractiveness ratings, effect size β_{dif} = -.14 for the objective attractiveness ratings).

The *Own Age* analysis applies to all participants. For this analysis, we examined the association of physical attractiveness with romantic desire only for faces who are within 10 years (older and younger) of the participant. A 40-year old participant might not consider a 21-year old to be a viable romantic partner, and by limiting the age range of the faces in the analysis, we could examine if the sex difference in the preference for physical attractiveness applies specifically to cases where participants rate partners who might conceivably be dating partners. Once again, these analyses did not substantively differ from the primary romantic desire analyses (i.e., effect size near zero for the participant-report attractiveness ratings, effect size β_{dif} = -.13 for the objective attractiveness ratings).

The *Unmarried participants* analysis applies specifically to participants who report that they are "Unmarried, but in a serious relationship" and "Single," and the *Single participants* analysis applies only to the participants who report that they are "Single." Many studies have found that single people evaluate partners more positively than individuals in a committed relationship (e.g., Simpson, Gangestad, & Lerma, 1990), and several theoretical perspectives in the close relationships tradition suggest that people in serious relationships might not be motivated to evaluate potential partners fairly and honestly (Lydon, 2010). Generally speaking, prior studies of sex differences in the association of physical attractiveness with romantic desire have tended not to document a meaningful role for relationship status (e.g., de Vries, 2010) or have not examined it as a moderator (Li et al., 2013; Townsend & Levy, 1990b). Nevertheless, the Unmarried and Single participants analyses examined if sex differentiated effects of physical attractiveness emerge among the participants who are most likely to be unbiased when evaluating potential partners. None of these exclusions affected the sex difference: Again, the effect size was near zero for the participant-report attractiveness ratings, but was β_{dif} = -.11 for both analyses using the objective attractiveness ratings.

The *Bottom 50th Percentile* and *Top 50th Percentile* analyses examined the association of physical attractiveness with romantic desire for faces rated in the bottom and top 50th percentile of attractiveness, respectively. The mate preference priority model (Li et al., 2013) suggests that the sex difference in the association of physical attractiveness with romantic desire is likely to be especially pronounced in the low-to-moderate range of attractiveness. These analyses examined this hypothesis. (The eligible faces for each analysis were determined on a participant-by-participant basis for the "participant report" tests; the eligible faces were determined for the whole sample based on the norming data for the "objective" test.) For the participant-report

attractiveness ratings, the sex differences trended in a direction opposite of the predictions generated by the mate preference priority model (i.e., the sex difference favored women for the bottom 50th percentile), but both the bottom and top 50th percentile analyses β_{dif} effect sizes were within the region of equivalence and are not large enough to merit substantive consideration. For the objective attractiveness ratings, the sex difference we detected in the primary analyses was (if anything) smaller when we restricted the dataset to bottom 50th percentile of targets (β_{dif} = -.08) than the overall analysis (β_{dif} = -.13); it was also smaller when we restricted the dataset to the top 50th percentile of targets (β_{dif} = -.07). In other words, the sex difference was strongest when the full range of targets was present. These results are inconsistent with a core postulate of the mate preference priority model (i.e., that the sex difference emerges especially strongly in the low-to-moderate range of attractiveness; Li & Meltzer, 2015; Li et al., 2013).

The *All Ps and photos < 35* analysis included all participants and photographs who are 35 years old or younger. Some perspectives suggest that samples in this age range are especially likely to reveal sex differences in the effects of physical attractiveness on romantic evaluations (Meltzer et al., 2014a, 2014b; Li & Meltzer, 2015). These analyses did not substantively differ from the primary romantic desire analyses (i.e., effect size near zero for the participant-report attractiveness ratings, effect size β_{dif} = -.14 for the objective attractiveness ratings). These results are inconsistent with the suggestion that participants in this age range are especially likely to reveal sex differences.

The *Same-sex desire* analyses examined the association of physical attractiveness with romantic desire for participants who reported that they are primarily attracted to same-sex partners. With respect to ideal partner preference ratings for attractiveness, biological sex has a stronger effect than sexual orientation: Gay men's ratings are comparable to heterosexual men's

ratings, both of which are higher than the ratings of heterosexual and lesbian women (Bailey et al., 1994; West, Popp, & Kenny, 2008). Therefore, the associations for gay men and lesbian women should presumably reveal a sex differentiated pattern that mirrors the heterosexual sample. The same-sex ratings are only included in this analysis (i.e., they are excluded from all earlier and subsequent analyses), and given that the sample of men and women attracted to same-sex partners is small, these effect sizes should be interpreted cautiously.

For the participant-report attractiveness ratings, no sex differences emerged, similar to the heterosexual sample. Surprisingly, the analyses using objective attractiveness ratings revealed a similar effect size for the sex difference (β_{dif} = .12) but in the opposite direction: attractiveness had a stronger effect on lesbian women's romantic desire than gay men's romantic desire. Despite the fact that gay men expressed a stronger preference for attractiveness than lesbian women (d = .52, see Methods section above), this sex difference significantly and substantively reversed when we examined the actual effect of objective attractiveness ratings on romantic desire.

Perceiver, target, and relationship effects. Our participants rated each target's attractiveness, and thus we were able to decompose the attractiveness measure into three components specified by the social relations model (Kenny, 1994, Kenny & La Voie, 1980). Specifically, for each attractiveness judgment, we could algebraically calculate a perceiver component (i.e., the extent to which the participant believed targets were attractive on average), a target component (i.e., the extent to which the target was consensually rated as attractive on average), and a relationship component (i.e., the extent to which a particular participant rated a particular target as attractive, above and beyond the relevant perceiver and target effect; see

Banchefsky, Westfall, Park, & Judd, 2016 for an illustration). In principle, sex differences could emerge on any of these three components. Thus, we tested the following model:

DV = attractiveness_target + attractiveness _perceiver + attractiveness _relationship + sex + attractiveness _target*sex + attractiveness _perceiver*sex + attractiveness _relationship*sex. (1)

We used the same covariance structure as the primary analyses above, and we subtracted the grand mean from each component (as in Joel, Eastwick, & Finkel, 2017). We conducted this regression equation three times: Once for the 2-item romantic desire dependent measure, once for the date DV item, and once for the serious relationship DV item (i.e., the first three rows of Table 1).

The results were more or less identical regardless of whether we examined romantic desire, the date item, or the serious relationship item as the dependent measure. First, the target component sex differences were significant and as large as the objective attractiveness effects in Table 1: romantic desire β_{dif} = -.14, 95% CI (-.14, -.13), t(360000) = -45.18, p < .001; date item β_{dif} = -.14, 95% CI (-.14, -.13), t(350000) = -50.49, p < .001, serious relationship item β_{dif} = -.12, 95% CI (-.13, -.12), t(350000) = -36.95, p < .001. Second, the perceiver effect sex differences were small and nonsignificant, much like the participant-report attractiveness analyses in Table 1: romantic desire β_{dif} = .00, 95% CI (-.02, .03), t(360000) = 0.31, p = .758; date item β_{dif} = .01, 95% CI (-.02, .04), t(350000) = 0.75, p = .456, serious relationship item β_{dif} = .00, 95% CI (-.03, .04), t(350000) = 0.26, p = .798. Third, the relationship effects were significant, small, and trended in the opposite direction of the target effects (i.e., attractiveness had a larger effect for women than men): romantic desire β_{dif} = .02, 95% CI (.02, .03), t(360000) = 18.83, p < .001; date

item β_{dif} = .02, 95% CI (.02, .03), t(350000) = 18.82, p < .001, serious relationship item β_{dif} = .03, 95% CI (.03, .03), t(350000) = 19.98, p < .001.

In summary, these social relations model component analyses revealed that (a) the extent to which the attractiveness level of photographs elicits romantic desire is stronger for female than male targets (i.e., target component effects), (b) there is no sex difference in the extent to which participants' general tendency to perceive attractiveness is associated with romantic desire (i.e., perceiver component effects), and (c) the extent to which a participant finds a target uniquely attractive is associated with romantic desire more strongly for women than for men (i.e., relationship component effects).

Level metric partner preference test. We also examined if people's own personal preference for physical attractiveness (based on the three *Stated Preference* items reported prior to the rating task) positively predicted the association of attractiveness with romantic desire (i.e., the preference × attractiveness interaction, also called the level metric; Eastwick et al., 2014a; Eastwick & Neff, 2012). Essentially, this interaction is the individual-differences analog of the sex difference test; it examines whether the slope of attractiveness predicting romantic desire is larger (i.e., more positive) for people who profess to have a strong preference for attractiveness than for people who profess a weak preference for attractiveness. This interaction follows theoretically from both evolutionary psychological perspectives (Li et al., 2013) and the ideals standards model (Simpson et al., 2001).

We tested this interaction using eight different permutations of our stated preference, attractiveness, and dependent variable measures (Table 3). The principle of compatibility (Ajzen & Fishbein, 1977) suggests that preferences may have greater predictive power when the preference and the dependent measure are equated for specificity. For this reason, the most

successful tests should presumably be revealed in the analyses that predict the date DV from the date preference (i.e., third and seventh rows in Table 3) and the serious relationship DV from the serious relationship preference (i.e., fourth and eighth rows in Table 3). All eight preference × attractiveness interactions were positive, significant, and fell inside the region of equivalence. In other words, the slope of attractiveness predicting romantic desire was stronger for participants with strong (vs. weak) stated preferences for attractiveness, but this effect was extremely small. The interaction β was approximately .03, and the effect sizes did not appear to be any stronger when the preference measure and dependent measure were (vs. were not) equated for specificity.

Exploratory Analyses

Stated-revealed preference correlations. Photo-rating tasks in prior studies (e.g., D. Wood & Brumbaugh, 2009) have obtained support for the level metric by conducting a variant of the above analysis. First, the researcher calculates a revealed preference (or "in-vivo" preference; Eastwick & Finkel, 2008) for each participant as the within-person slope (i.e., personal regression β) between the attractiveness measure (i.e., second column in Table 3) and the dependent measure (i.e., third column in Table 3) across all the photographs that he/she rated. Second, this revealed preference is correlated (across all participants) with the stated preference. This analysis (which was omitted from the preregistration) is presented in the rightmost column in Table 3. In all cases, the stated-revealed preference correlation was significant and a small-to-moderate effect size (average r = ~.23), which approximates the average estimate obtained by D. Wood and Brumbaugh (2009) on a similar photo-rating task (i.e., r = ~.18). In other words, whereas the preference × attractiveness interactions were extremely small, the conceptually identical stated-revealed preference correlation was substantial. We return to explore possible

reasons for this striking disparity—and potential implications for best practices in measurement—in the Discussion.

Li and Meltzer's (2015) multi-moderation prediction. We conducted an additional, exploratory analysis to test Li and Meltzer's (2015) specific prediction that the largest sex difference should emerge when researchers "take into consideration simultaneously" (p. 95) the following four conditions: (a) only the serious relationship dependent measure, (b) only the objective measure of attractiveness, (c) only targets in the bottom 50^{th} percentile, and (d) only participants and targets under age 35. In this analysis, the sex difference remained unchanged from the overall analysis: β_{dif} = -.09, 95% CI(-.16,-.04), t(92000) = -3.66, p < .001; β_{Men} = .21, β_{Women} = .11. These results are inconsistent with the claim that the largest sex difference will emerge under the complex confluence of these four specific conditions.

Discussion

Men reliably indicate on rating scales that they prefer attractiveness in a romantic partner more than women do (Buss, 1989). Nevertheless, the literature examining the potential downstream consequences of this sex difference is filled with contradictory findings. This registered report, which was peer reviewed and approved prior to data collection, sheds light on this issue by testing whether there are sex differences in the appeal of attractiveness in a photograph-rating context. Although the photograph-rating context is one in which nearly all scholars (including sex-differences skeptics) have posited the existence of a sex difference (Eastwick et al., 2014a), prior studies on this topic do not meet modern standards of statistical power because they tended to sample a very small number of targets. The current study used a large sample of participants and targets and documented several findings with important implications for theory development in the mate preferences literature.

Key Findings

Objective measures of attractiveness reveal the sex difference. First, the sex difference in the association of attractiveness with romantic desire emerged consistently when using objective (i.e., independent raters') assessments of attractiveness. In contrast, the sex difference did not emerge when participants themselves evaluated the target's attractiveness (see also Tshkay, Clout, & Rule, 2017). In addition, when we separated participants' attractiveness evaluations into separate social relations model components (i.e., target, perceiver, and relationship effects), a sex difference in the male direction emerged only for the target component of attractiveness—the component that represents the consensus attractiveness score for each target and is conceptually analogous to the independent raters' judgments. Together, these primary analysis findings are consistent with prior suggestions that sex differences in the appeal of attractiveness may be more likely to emerge when using objective (e.g., consensus) measures of attractiveness rather than participants' personal evaluations of attractiveness (Li & Meltzer, 2015).

All remaining moderators were unsupported. Second, none of the other moderators affected the size of the sex difference in the predicted manner. In the existing literature on sex differences in the appeal of attractiveness, several of these moderators have been invoked to explain cross-study variability in sex difference effect size estimates. In other words, when the presence or size of sex differences appears to vary from one study to the next, scholars have often proposed moderators to explain those fluctuations (rather than assuming that these fluctuations are due to sampling variability). For example, some scholars have speculated that sex differences will be larger when participants (a) consider long-term serious relationships, (b)

are under age 35, and (c) evaluate targets who are in the low-to-moderate range of physical attractiveness (Li & Meltzer, 2015; Li et al., 2013; Meltzer et al., 2014a, 2014b; Schmitt, 2014).

In the current study, we conducted the first empirical tests of these proposed moderators. For the "serious relationship" and "all Ps and photos < 35" tests, the size of the sex difference was identical to the primary analysis test; that is, these moderators made no difference. When we restricted analyses to the low-to-moderate range of attractiveness in the "bottom 50th percentile" tests, the sex difference actually decreased (in the case of objective ratings) or reversed (in the case of own ratings), rather than increasing (Li et al., 2013). Thus, there is no empirical evidence to support the notion that variables such as "only long-term, serious relationships," "only participants under 35," and "only the low-to-moderate range of attractiveness" moderate the size of the sex difference assessed here. Given these findings, sampling fluctuation across often underpowered studies remains the most plausible explanation for why some studies show sex differences and others do not (see also Eastwick et al., 2014a, 2014b; Eastwick, Neff, et al., 2014). We recommend that going forward, scholars prioritize highly powered studies, registered reports, and meta-analytic tools to further probe these questions.

The sex difference is a property of the targets. Third, the sex difference in the association of attractiveness with romantic desire that we documented in this study seems to largely reflect the properties of the targets (i.e., faces) that are being rated rather than properties of the people making the ratings. That is, it is accurate to say that women's attractiveness (as depicted in photographs) elicits more romantic desire than men's attractiveness does; it would not be accurate to say that men experience a stronger link between attractiveness and romantic desire than women do. Importantly, this distinction underscores why the random sampling of targets may be as crucial for claims about generalizability as the random sampling of participants

in this domain. Thus, designs that probe for sex differences with small numbers of targets (or that attempt to handpick a handful of attractive and unattractive confederates) cannot offer strong evidence for or against the sex difference, even if they have large samples of raters; small numbers of targets will create a bouncy, unstable sex difference estimate. Studies in which raters also serve as targets are likely to offer especially promising, appropriately powered tests of sex differences (e.g., couples designs and speed-dating designs).

Gay and lesbian participants reveal a preference reversal. Fourth, the importance of target sampling is also well illustrated by the same-sex rating analyses. These analyses revealed a never-before-seen mate preference reversal: Although gay men expressed a stronger stated preference for attractiveness than lesbian women, the attractiveness of the photographs predicted lesbian women's romantic desire more strongly than it predicted gay men's romantic desire. Importantly, the objective attractiveness effect for gay men (r = .26) was nearly identical to the overall effect for straight women (r = .28)—the two groups who rated the male photographs. Also, the objective attractiveness effect for lesbian women (r = .38) was nearly identical to the overall effect for straight men (r = .41)—the two groups who rated the female photographs. The sample of individuals attracted to same-sex partners was small, and so the effect sizes for this group should be considered provisional. Nevertheless, the fact that the sex difference for the gay/lesbian stated preference is the opposite of the sex difference in the gay/lesbian attractiveness-desire association offers some of the most intriguing evidence to date that stated and revealed preferences operate through very distinct mechanisms; the process of translating a revealed preference into a stated preference may be imperfect and subject to various cognitive and/or motivational biases (e.g., Smith, Eastwick, & Ledgerwood, 2017). Perspectives that downplay the psychological importance of the distinction between stated and revealed

preferences (e.g., Gerlach, Arslan, Schultze, Reinhard, & Penke, in press) may struggle to explain dissociations like the one documented here.

The two level metric approaches are not (statistically) identical. Fifth and finally, the level metric tests contained some important analytic lessons. These tests examined whether people who exhibited a stronger association between attractiveness and romantic desire (i.e., a stronger revealed preference; D. Wood & Brumbaugh, 2009) also had stronger stated preferences for attractiveness. The preregistered stated preference × attractiveness interactions that we conducted to test this hypothesis were significant and in the predicted (positive) direction, but they were extremely small (β = .03 on average; Table 3). In contrast, when we calculated each participant's own personal revealed preference and then correlated these estimates with stated preferences, the effect sizes were moderate in magnitude (r = .23 on average).

Why do these two statistical approaches—both of which have been used in past research on stated and revealed preferences (Eastwick & Finkel, 2008; Li et al., 2013; D. Wood & Brumbaugh, 2009)—reveal such different effect sizes if they test the same phenomenon? One likely explanation is measurement reliability. Specifically, each photograph rating can be conceptualized as a highly unreliable measure of a given participant's revealed preference for attractiveness. Aggregating across these ratings provides a much more reliable measure. In other words, a participant's revealed preference for attractiveness becomes clarified across her 300 ratings (like a 300-item scale) and hence correlates moderately strongly with her stated preference, but her stated preference is negligibly related to the extent to which a particular target's attractiveness is romantically inspiring (D. Wood, personal communication, November 6, 2017).

We draw two important conclusions from the disparate results produced by these two conceptually similar measures. First, tests of the stated preference × attractiveness interaction (e.g., Eastwick, Eagly, Finkel, & Johnson, 2011; Li et al., 2013) rely on a highly unreliable measure and are therefore likely to be seriously underpowered, whereas tests of stated-revealed preference correlations may be more robust (e.g., Eastwick & Finkel, 2008; D. Wood & Brumbaugh, 2009); scholars should accord greater confidence to conclusions drawn from the latter measure in the existing literature. Second, going forward, we strongly recommend that scholars employ the more reliable measure when possible (stated-revealed preference correlations).

Strengths, Limitations, and Future Directions

This study had a number of strengths. We achieved two important goals with our preregistration. First, preregistering our analysis plan allowed us to avoid researcher degrees of freedom and the resulting inflation of false positives (Ledgerwood, Soderberg, & Sparks, 2017; Sagarin, Ambler, & Lee, 2014). Second, because we preregistered directional hypothesis tests that follow from particular theoretical perspectives, we were able to interpret the results of those tests vis-à-vis theory falsification (e.g., the fact that the sex difference did not increase in the low-to-moderate range of attractiveness falsifies a postulate of the mate preference priority model; Li et al., 2013). Also, we used a large sample of both participants and targets, which means that we limited the likelihood of Type II error, and also that our effect size estimates are relatively precise. Finally, by sampling our targets from a naturalistic set of male and female photographs, we can be confident that our findings should generalize to the range of attractiveness that characterizes typical male and female populations.

Of course, this study also had several limitations. First and foremost, it only examined sex differences in a photograph-rating context; according to past theoretical and empirical work (Eastwick et al., 2014a), the sex difference was especially likely to emerge on this relatively low-complexity task. This study does not address the effect size of the sex difference in association between objective attractiveness and romantic desire in initial attraction contexts or established relationship contexts. (In meta-analyses, both sex differences are approximately $q = .05$ or smaller and not significant; Eastwick et al., 2014a; Eastwick, Neff et al., 2014; registered reports seem like an especially useful next step for researchers interested in further investigating the possibility of sex differences in this context). Secondly, the same-sex findings deserve replication and additional scrutiny—especially given the small sample size of this group—before we conclude that the association of attractiveness with romantic desire is sex-differentiated *in the opposite direction* of the stated preference sex difference among these individuals. Although a number of studies have found that gay men claim to desire attractiveness in a partner more than lesbian women (i.e., stated preferences; Bailey et al., 1994; West et al., 2008), we know of only one other study that examined the association between photograph-attractiveness and romantic desire among gay men and lesbians (Bailey, Kim, Hills, & Linsenmeier, 1997, Study 3), and it found no sex difference ($B = .44$ for gay men, $B = .46$ for lesbian women). Additional, preregistered research should examine how strongly various attributes are associated with romantic desire for gay men and lesbian women, both in photograph-rating and face-to-face dating contexts.

Conclusion

Physical attractiveness is one of the most extensively studied constructs in social psychology (Langlois et al., 2000), and scientists will surely continue to debate the extent to

which the association of physical attractiveness with romantic evaluations is sex differentiated. The current study is one of the strongest demonstrations to date that this association is stronger for heterosexual men than for women in photograph-rating contexts using objective ratings of physical attractiveness. Given that the study was a highly powered registered report using many participants and targets, it offers a useful benchmark for the effect size of this sex difference (β_{dif} = ~.13; approximately equivalent to q = .13, a small effect size).

From a theoretical perspective, the current study also makes a number of useful contributions. For example, these findings are consistent with models suggesting that the sex difference is present in hypothetical/photograph-rating contexts but absent in initial attraction and close relationships contexts because the former class of contexts facilitates the use of abstract sources of information, such as stated preferences (Eastwick et al., 2011, 2014a). Also, this study afforded the opportunity to test a variety of moderators—moderators that follow from specific theoretical perspectives that purportedly explain cross-study variability in the presence vs. absence of the attractiveness sex difference (Li & Meltzer, 2015; Meltzer et al., 2014a, 2014b). The results were consistent with one of the proposed moderator predictions (e.g., sex differences were stronger when using objective ratings of attractiveness); they were inconsistent with the other moderator predictions (e.g., sex differences were not stronger in serious relationship contexts, in the low-to-moderate range of attractiveness, etc.). In summary, the current study demonstrates how registered reports and detailed preregistered analysis plans can improve our understanding of the magnitude of sex differences and advance debates in the psychological sciences.

References

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

Bailey, J. M., Gaulin, S., Agyei, Y., & Gladue, B. A. (1994). Effects of gender and sexual orientation on evolutionarily relevant aspects of human mating psychology. *Journal of Personality and Social Psychology, 66*, 1081-1093.

Bailey, J. M., Kim, P. Y., Hills, A., & Linsenmeier, J. A. (1997). Butch, femme, or straight acting? Partner preferences of gay men and lesbians. *Journal of Personality and Social Psychology, 73*, 960-973.

Banchefsky, S., Westfall, J., Park, B., & Judd, C. M. (2016). But you don't look like a scientist!: Women scientists with feminine appearance are deemed less likely to be scientists. *Sex Roles, 75*, 95-109.

Brown-Iannuzzi, J.L., Dotsch, R., Cooley, E., & Payne, B.K. (in press). The relationship between mental representations of welfare recipients and attitudes toward welfare. *Psychological Science*.

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

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review, 100*, 204-232.

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.

Buunk, B. P., Dijkstra, P., Fetchenhauer, D., & Kenrick, D. T. (2002). Age and gender differences in mate selection criteria for various involvement levels. *Personal Relationships, 9,* 271-278.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Psychology Press.

Coffman, K. B., Coffman, L. C., & Ericson, K. M. M. (2016). The size of the LGBT population and the magnitude of antigay sentiment are substantially underestimated. *Management Science, 63,* 3168-3186.

Cunningham, M. R. (1986). Measuring the physical in physical attractiveness: Quasi-experiments on the sociobiology of female facial beauty. *Journal of personality and social psychology, 50,* 925-935.

Cunningham, M. R., Barbee, A. P., & Pike, C. L. (1990). What do women want? Facialmetric assessment of multiple motives in the perception of male facial physical attractiveness. *Journal of personality and social psychology, 59,* 61-72.

de Vries, J. M. A. (2010). Impact of self-descriptions and photographs on mediated dating interest. *Marriage & Family Review, 46,* 538-562.

de Vries, J. M. A., Swenson, L., & Walsh, R. P. (2007). Hot picture or great self-description: Predicting mediated dating success with parental investment theory. *Marriage and Family Review, 42,* 7-34.

Eastwick, P. W., Eagly, A. H., Finkel, E. J., & Johnson, S. E. (2011). Implicit and explicit preferences for physical attractiveness in a romantic partner: A double dissociation in predictive validity. *Journal of Personality and Social Psychology, 101,* 993-1011.

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

Eastwick, P. W., Finkel, E. J., & Eagly, A. H. (2011). When and why do ideal partner preferences affect the process of initiating and maintaining romantic relationships? *Journal of Personality and Social Psychology, 101*, 1012-1032. doi: 10.1037/a0024062

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

Eastwick, P. W., Luchies, L. B., Finkel, E. J., & Hunt, L. L. (2014b). The many voices of Darwin's descendants: Reply to Schmitt (2014). *Psychological Bulletin, 140*, 673-681.

Eastwick, P. W., & Neff, L. A. (2012). Do ideal partner preferences predict divorce? A tale of two metrics. *Social Psychological and Personality Science, 3*, 667-674.

Eastwick, P. W., Neff, L. A., Luchies, L. B., Finkel, E. J., & Hunt, L. L. (2014). Is a meta-analysis a foundation or just another brick? Comment on Meltzer, McNulty, Jackson, & Karney (2014). *Journal of Personality and Social Psychology, 106*, 429-434.

Feingold, A. (1990). Gender differences in effects of physical attractiveness on romantic attraction: A comparison across five research paradigms. *Journal of Personality and Social Psychology, 59*, 981-993.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (in press). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*.

Fletcher, G. J. O., Kerr, P. S., Li, N. P., & Valentine, K. A. (2014). Predicting romantic interest and decisions in the very early stages of mate selection standards, accuracy, and sex differences. *Personality and Social Psychology Bulletin, 40*, 540-550.

Gerlach, T. M., Arslan, R. C., Schultze, T., Reinhard, S. K., & Penke, L. (in press). Predictive validity and adjustment of ideal partner preferences across the transition into romantic relationships. *Journal of Personality and Social Psychology*.

Hauser, D. J., & Schwarz, N. (2016). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods, 48*, 400-407.

Hill, R. (1945). Campus values in mate-selection. *Journal of Home Economics, 37*, 554-558.

Hitsch, G. J., Hotacsu, A., & Ariely, D. (2010). What makes you click? Mate preferences in online dating. *Quantitative Marketing and Economics*, *8*, 393-427.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine, 2*, e124.

Joel, S., Eastwick, P. W., & Finkel, E. J. (2017). Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychological Science, 28*, 1478-1489.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of personality and social psychology, 103*, 54-69.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*, 601-625.

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

Kenny, D. A., & La Voie, L. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 18, pp. 141-182). New York: Academic Press.

Kenrick, D. T., Groth, G. E., Trost, M. R., & Sadalla, E. K. (1993). Integrating evolutionary and social exchange perspectives on relationships: Effects of gender, self-appraisal, and involvement level on mate selection criteria. *Journal of Personality and Social Psychology, 64*, 951-969.

Kenrick, D. T., Sadalla, E. K., Groth, G., & Trost, M. R. (1990). Evolution, traits, and the stages of human courtship: Qualifying the parental investment model. *Journal of Personality, 58*, 97-116.

Kiernan, K., Tao, J., & Gibbs, P. (2012). *Tips and strategies for mixed modeling with SAS/STAT procedures.* SAS Global Forum, paper 332-2012.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability. *Social psychology, 45*, 142-152.

Lakens, D. (in press). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*.

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.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*, 390-423.

Ledgerwood, A., Soderberg, C. K., & Sparks, J. (2017). Designing a study to maximize informational value. In J. Plucker & M. Makel (Eds.), *Toward a more perfect psychology: Improving trust, accuracy, and transparency in research* (pp. 33-58). Washington, DC: American Psychological Association.

Lee, L., Loewenstein, G., Ariely, D., Hong, J., & Young, J. (2008). If i'm not hot, are you hot or not? Physical-attractiveness evaluations and dating preferences as a function of one's own attractiveness. *Psychological Science, 19*, 669-677.

Lewandowski, G. W., Aron, A., & Gee, J. (2007). Personality goes a long way: The malleability of opposite-sex physical attractiveness. *Personal Relationships, 14*, 571-585.

Li, N. P., & Kenrick, D. T. (2006). Sex similarities and differences in preferences for short-term mates: What, whether, and why. *Journal of Personality and Social Psychology, 90*, 468-489.

Li, N. P., & Meltzer, A. L. (2015). The validity of sex-differentiated mate preferences: Reconciling the seemingly conflicting evidence. *Evolutionary Behavioral Sciences, 9*, 89-106.

Li, N. P., Yong, J. C., Tov, W., Sng, O., Fletcher, G. J. O., Valentine, K. A., . . . Balliet, D. (2013). Mate preferences do predict attraction and choices in the early stages of mate selection. *Journal of Personality and Social Psychology, 105*, 757-776. doi: 10.1037/a0033777

Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: Evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*, 1638-1659.

Lydon, J. E. (2010). How to forego forbidden fruit: The regulation of attractive alternatives as a commitment mechanism. *Social and Personality Psychology Compass, 4*, 635-644.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods, 47*, 1122-1135.

Meltzer, A. L., McNulty, J. K., Jackson, G. L., & Karney, B. R. (2014a). Sex differences in the implications of partner physical attractiveness for the trajectory of marital satisfaction. *Journal of Personality and Social Psychology, 106*, 418-428.

Meltzer, A. L., McNulty, J. K., Jackson, G. L., & Karney, B. R. (2014b). Men still value physical attractiveness in a long-term mate more than women: Rejoinder to Eastwick, Neff, Finkel, Luchies, and Hunt (2014). *Journal of Personality and Social Psychology, 106*, 435-440.

Montoya, R. M. (2008). I'm hot, so i'd say you're not: The influence of objective physical attractiveness on mate selection. *Personality and Social Psychology Bulletin, 34*, 1315-1331.

Olderbak, S. G., Malter, F., Wolf, P. S. A., Jones, D. N., & Figueredo, A. J. (2017). Predicting romantic interest at zero acquaintance: Evidence of sex differences in trait perception but not in predictors of interest. *European Journal of Personality, 31*, 42–62.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*, 11087-11092.

Perrett, D. I., Lee, K., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D., . . . Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature, 394*, 884-887.

Regan, P. C., Levin, L., Sprecher, S., Christopher, F. S., & Cate, R. (2000). Partner preferences: What characteristics do men and women desire in their short-term sexual and long-term romantic partners? *Journal of Psychology & Human Sexuality, 12*, 1-21.

Ritter, S. M., Karremans, J. C., & van Schie, H. T. (2010). The role of self-regulation in derogating attractive alternatives. *Journal of Experimental Social Psychology, 46*, 631-637.

Rudder, C. (2010, January 20). The 4 big myths of profile pictures. Retrieved from https://theblog.okcupid.com/the-4-big-myths-of-profile-pictures-41bedf26e4d#.uv11b6ohg

Rudder, C. (2014). *Dataclysm: Who we are (when we think no one's looking)*: Random House Canada.

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

Schmitt, D. P. (2014). On the proper functions of human mate preference adaptations: Comment on Eastwick, Luchies, Finkel, and Hunt (2014). *Psychological Bulletin, 140*, 666-672.

Simonsohn, U. (2014). [17] No-way interactions. Retrieved from http://web.archive.org/web/20150206205257/http://data-colada.org/2014/03/12/17-no-way-interactions-2/

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

Simpson, J. A., Gangestad, S. W., & Lerma, M. (1990). Perception of physical attractiveness: Mechanisms involved in the maintenance of romantic relationships. *Journal of Personality and Social Psychology, 59*, 1192-1201.

Smith, L. K., Eastwick, P. W., & Ledgerwood, A. (2017). *How do people infer preferences for attributes?* Unpublished Manuscript, University of California, Davis, CA.

Sprecher, S., Sullivan, Q., & Hatfield, E. (1994). Mate selection preferences: Gender differences examined in a national sample. *Journal of Personality and Social Psychology, 66*, 1074-1080.

Tskhay, K. O., Clout, J. M., & Rule, N. O. (2017). The impact of health, wealth, and attractiveness on romantic evaluation from photographs of faces. *Archives of Sexual Behavior*, *46*, 2365–2376.

Townsend, J. M., & Levy, G. D. (1990). Effects of potential partners' physical attractiveness and socioeconomic status on sexuality and partner selection. *Archives of Sexual Behavior, 19*, 149-164.

Townsend, J. M., & Roberts, L. W. (1993). Gender differences in mate preference among law students: Divergence and convergence of criteria. *The Journal of Psychology*, *127*, 507-528.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25*, 1115-1125.

Wenzel, A., & Emerson, T. (2009). Mate selection in socially anxious and nonanxious individuals. *Journal of Social and Clinical Psychology*, *28*, 341-363.

West, T. V., Popp, D., & Kenny, D. A. (2008). A guide for the estimation of gender and sexual orientation effects in dyadic data: An actor-partner interdependence model approach. *Personality and Social Psychology Bulletin, 34*, 321-336.

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science, 10*, 390-399.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*, 2020-2045.

Wood, D., & Brumbaugh, C. C. (2009). Using revealed mate preferences to evaluate market force and differential preference explanations for mate selection. *Journal of Personality and Social Psychology, 96*, 1226-1244.

Wood, W., & Eagly, A. H. (2002). A cross-cultural analysis of the behavior of women and men: Implications for the origins of sex differences. *Psychological Bulletin, 128*, 699-727.

Author Note

Paul W. Eastwick and Leigh K. Smith, Department of Psychology, University of California, Davis. We wish to thank Alison Ledgerwood for providing feedback on earlier drafts of this manuscript. Correspondence concerning this article should be addressed to Paul Eastwick, University of California, Davis, Department of Psychology, Young Hall, Davis, CA 95616. E-mail may be sent to eastwick@ucdavis.edu.

Table 1 – Full Sample Analyses

	Attractiveness effect size								Sample size			
	Men	Women	Test of Sex difference						Men		Women	
Analysis	β	β	β_{dif}	t	95% CI	90% CI	Sig	Equ	Part. N	Stim. n	Part. N	Stim. n
Participant report												
Romantic desire	.81***	.81***	-.00	-0.24	[-.03, .02]	[-.03, .02]	N	Y	609	305	595	288
Date item	.83***	.83***	-.01	-0.69	[-.03, .02]	[-.03, .01]	N	Y	609	305	595	288
Serious relationship item	.77***	.78***	.01	0.60	[-.02, .04]	[-.02, .04]	N	Y	609	305	595	288
Objective												
Romantic desire	.41***	.28***	-.13	-5.19	[-.18, -.08]	[-.17, -.09]	Y	N	609	305	595	288
Date item	.42***	.28***	-.13	-5.32	[-.18, -.08]	[-.18, -.09]	Y	N	609	305	595	288
Serious relationship item	.39***	.27***	-.12	-4.86	[-.17, -.07]	[-.16, -.08]	Y	N	609	305	595	288

Note: Participant *N*s are estimated after accounting for exclusions. βs are calculated on standardized stimuli and can be interpreted similarly to effect size *r*, and β_{dif} can be interpreted similarity to effect size *q*. Sig column indicates whether the 95% CI for the sex difference does (N) or does not (Y) include zero; Equ column indicates whether the 90% CI for the sex difference does (Y) or does not (N) reside within the region of equivalence (-.10, .10).

Table 2 – Auxiliary Subsample Analyses

	Attractiveness effect size								Sample size			
	Men	Women	Test of Sex difference						Men		Women	
Analysis	β	β	β_{dif}	t	95% CI	90% CI	Sig	Equ	Part. N	Stim. n	Part. N	Stim. n
Participant report												
Own race	.83***	.81***	-.02	-1.32	[-.05, .01]	[-.04, .00]	N	Y	581	305	563	288
Similar age	.81***	.82***	.00	0.03	[-.03, .03]	[-.02, .02]	N	Y	609	305	595	288
Unmarried participants	.81***	.82***	.01	0.30	[-.03, .04]	[-.02, .04]	N	Y	397	305	288	288
Single participants	.80***	.81***	.01	0.36	[-.04, .06]	[-.03, .05]	N	Y	276	305	158	288
Bottom 50th percentile	.72***	.78***	.05	2.16	[.00, .09]	[.01, .09]	Y	Y	609	305	595	288
Top 50th percentile	.76***	.73***	-.04	-3.04	[-.06, -.01]	[-.06, -.02]	Y	Y	609	305	595	288
All Ps and photos < 35	.81***	.82***	.00	0.00	[-.03, .03]	[-.03, .03]	N	Y	440	265	392	230
Same-sex desire	.80***	.79***	-.01	-0.13	[-.12, .10]	[-.10, .08]	N	Y	46	288	57	305
Objective												
Own race	.45***	.30***	-.14	-5.64	[-.19, -.09]	[-.19, -.10]	Y	N	581	305	563	288
Similar age	.42***	.29***	-.13	-5.16	[-.18, -.08]	[-.17, -.09]	Y	N	609	305	595	288
Unmarried participants	.40***	.28***	-.11	-4.27	[-.17, -.06]	[-.16, -.07]	Y	N	397	305	288	288
Single participants	.38***	.27***	-.11	-3.61	[-.16, -.04]	[-.15, -.06]	Y	N	250	305	250	288
Bottom 50th percentile	.21***	.13***	-.08	-3.03	[-.13, -.03]	[-.12, -.04]	Y	N	609	153	595	146
Top 50th percentile	.27***	.20***	-.07	-2.08	[-.14, -.00]	[-.13, -.02]	Y	N	609	152	595	142
All Ps and photos < 35	.40***	.26***	-.14	-4.96	[-.19, -.08]	[-.18, -.09]	Y	N	440	265	392	230
Same-sex desire	.26***	.38***	.12	2.54	[.03, .20]	[.04, .19]	Y	N	46	288	57	305

Note: Participant Ns are estimated after accounting for exclusions. βs are calculated on standardized stimuli and can be interpreted similarly to effect size r, and β_{dif} can be interpreted similarity to effect size q. Sig column indicates whether the 95% CI for the sex difference does (N) or does not (Y) include zero; Equ column indicates whether the 90% CI for the sex difference does (Y) or does not (N) reside within the region of equivalence (-.10, .10).

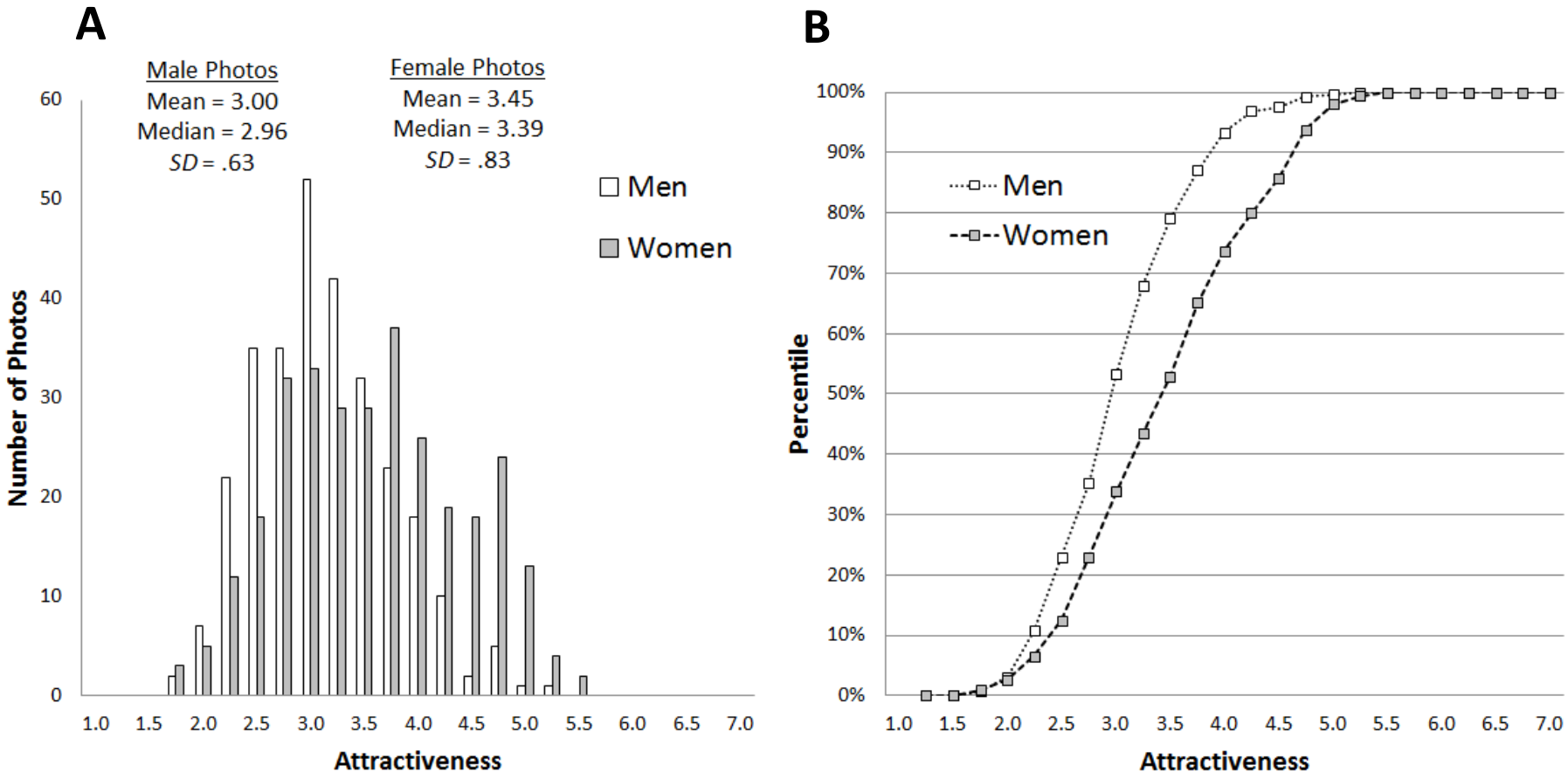
Table 3 – Level Metric Ideal Partner Preference Tests

Stated Preference measure		Attractiveness measure		Dependent measure	Interaction β	t	95% CI	90% CI	Sig	Equ	Stated-Revealed $r^{\dagger}$
3-item average	×	Participant report	=	Romantic desire	.03	4.48	[.02, .04]	[.02, .04]	Y	Y	.27***
Ideal item	×	Participant report	=	Romantic desire	.02	3.44	[.01, .04]	[.01, .03]	Y	Y	.25***
Date item	×	Participant report	=	Date item	.03	5.63	[.02, .05]	[.02, .04]	Y	Y	.27***
Serious item	×	Participant report	=	Serious item	.03	4.33	[.02, .05]	[.02, .05]	Y	Y	.22***
3-item average	×	Objective	=	Romantic desire	.03	5.27	[.02, .04]	[.02, .04]	Y	Y	.20***
Ideal item	×	Objective	=	Romantic desire	.02	3.95	[.01, .04]	[.01, .03]	Y	Y	.19***
Date item	×	Objective	=	Date item	.04	7.31	[.03, .05]	[.03, .05]	Y	Y	.24***
Serious item	×	Objective	=	Serious item	.03	4.07	[.01, .04]	[.02, .04]	Y	Y	.16***

Note: Sig column indicates whether the 95% CI for the Preference × Attractiveness interaction does (N) or does not (Y) include zero; Equ column indicates whether the 90% CI for the Preference × Attractiveness interaction does (Y) or does not (N) reside within the region of equivalence (-.10, .10). These analyses used measures that were standardized across the entire sample—not within-sex like the sex differences analyses—although interaction βs calculated using measures standardized within sex revealed nearly identical effect sizes. The revealed preference is each participant's personal regression β predicting the dependent measure (third column) from the attractiveness measure (second column).

† = column omitted from preregistered analysis plan

Figure 1 – Norming Data on Attractiveness in the Chicago Face Database



Attractiveness ratings (on a 1-7 scale) of the 288 male photos and 305 female photos under age 50 in the Chicago face database plotted as a histogram (Panel A) and by percentile (Panel B). Men and women cannot meaningfully be matched on numerical ratings because the distributions of men and women differ substantially at many points along the attractiveness continuum (e.g., a 3.0 on attractiveness is the 34th percentile for women but the 53rd percentile for men; see Panel B).