Social Cognition, Vol. 23, No. 3, 2005, pp. 257-278

CORRELATED AVERAGES VS. AVERAGED CORRELATIONS: DEMONSTRATING THE WARM GLOW HEURISTIC BEYOND AGGREGATION

Benoît Monin
Stanford University

Daniel M. Oppenheimer
Princeton University

Three studies demonstrate the warm glow heuristic (Monin, 2003) without relying on aggregated ratings, and illustrate the important distinction between correlating average ratings versus averaging individual correlations. In Study 1, we re–analyze previous data correlating individual ratings with aggregates from another small sample of raters. In Study 2, we correlate individual familiarity ratings with normed attractiveness from a large sample of raters ($n > 2{,}500$). Study 3 bypasses the issue of aggregates altogether by having participants provide both attractiveness and familiarity ratings and computing correlations within participants. Despite this more conservative approach, the results of all three studies support the existence of the beautiful–is–familiar phenomenon.

It is great to be good–looking. The mere fact of being physically attractive apparently improves one's life outcomes significantly. For example, attractive people are more likely to be helped (Chaiken, 1979), less likely to get punished (Downs & Lyons, 1991), and tend to earn more money (Hamermesh & Biddle, 1994). People make all kinds of positive inferences based on attractive-

The second author was supported by a National Science Foundation graduate research fellowship. We would like to thank Peter Finlayson for his help in collecting data for Study 2, and Dan Yarlett for his help in collecting data for Study 3.

Address correspondence to Benoît Monin, Department of Psychology, Jordan Hall, Stanford University, Stanford, CA 94305; E-mail: monin@stanford.edu.

ness alone: Attractive people seem more intelligent, more successful, more socially skilled, better adjusted, and in general are thought to possess more desirable qualities (Dion, Berscheid, & Walster, 1972; Eagly, Ashmore, Makhijani, & Longo, 1991).

Recently, it was discovered that the impact of attractiveness goes beyond personality inferences, and can actually influence whether we think we have seen a face before: Attractive people seem more familiar, and are more likely to be recognized, even on first encounter (Vokey & Read, 1992; Monin, 2003). In line with recent findings suggesting that positivity can cue familiarity (Garcia–Marques, Mackie, Claypool, & Garcia–Marques, 2004), Monin attributed this beautiful–is–familiar effect to a "warm glow heuristic" in which people use positive affective reactions to stimuli to infer familiarity (Monin, 2003; Corneille, Monin, & Pleyers, in press). In support of this interpretation, not only do attractive faces look familiar, but positive words are also more likely to seem familiar than neutral or negative words. This interplay between cognition and affect was foreshadowed by Zajonc (1980) when he proposed that our first reaction to stimuli is affective and that this first reaction colors subsequent judgments.

However, some demonstrations of the phenomenon exhibit a possible shortcoming. For example, Monin (2003, Study 1) showed a set of photographs to two groups of participants and demonstrated that the average rating of familiarity by a group of 40 judges correlated highly with the average rating of attractiveness by another group of 34 judges ($r = .64$). Subsequent studies used more sophisticated techniques, but retained the feature that average ratings on a first dimension by one group are correlated with average ratings on a second dimension by another group. This approach provides good evidence that if a given picture is rated as attractive, it is likely to be rated as familiar. However, correlating these group aggregates falls short of demonstrating the process assumed to underlie the warm glow heuristic at the level of the individual: It was assumed that if a picture is attractive for a given participant, it should also be more familiar to him or her. Using the picture as the unit of analysis does not afford the opportunity to test this phenomenon at the individual level as it tests the correlation of averages, rather than looking at averages of correlations.

AVERAGE CORRELATIONS VERSUS CORRELATION OF AVERAGES

The difference between correlated averages and averaged correlations is easily overlooked, and yet the two can differ widely. The example in Table 1 dramatically illustrates the fallacy of equating the two statistics. Imagine that two judges rate each of four stimuli, a, b, c, and d, on two separate dimensions A and B. Judge 1, for instance, rates stimulus b as a 2 on dimension A. When correlations are computed within judges, the two dimensions are correlated negatively for both judges, and the average correlation is strongly negative ($r = -.80$). However, when we start out by computing an average across judges for each stimulus, this correlation of averages is strongly positively correlated ($r = +1.00$). This extreme example demonstrates the important difference between averaging correlations and correlating averages. It would be accurate (based on the averaging method) to claim that the higher a stimulus is on dimension A (on average), the higher it will be (on average) on dimension B. It would not be accurate, however, to claim that this correspondence is maintained at the individual level, or to develop a theory of mental processes that posits that people use A when making judgments of B.[1]

The issue is not a new one (e.g., Zajonc, 1962; Gordon, 1924), but it was revived recently by Nickerson (1995), who criticized Kahneman and Knetsch's (1992) studies correlating willingness to pay (WTP) to address a particular social issue and moral outrage. Using a design akin in logic to Monin's (2003), Kahneman and Knetsch obtained WTP from one group of judges and moral outrage from another group. When they correlated the mean moral outrage for each issue and the median willingness to pay for the same issue, they found correlations as high as +.77. Although this does test the prediction that a given issue is more likely to elicit high WTP if it also elicits

1. The discrepancy observed in Table 1 between aggregated correlation and correlated aggregation is exacerbated because of the low agreement between the two judges ($r = -.80$ for both dimensions). When reliability between judges is assessed to be high enough (Rosenthal, 1987), it is quite legitimate to rely on average ratings to estimate a criterion. For example, Ambady & Rosenthal (1993) correlated average personality ratings based on 30–second silent clips of graduate students teaching by nine judges with average end–of–the–semester student ratings, and found a correlation of .76, $p < .001$.

TABLE 1. Simulation Demonstrating the Possible Disjunction between a Correlation of Averages, *r* (CA) = +1.00, and an Average of Correlations, *r* (AC) = –.80.

		Stimulus				
		a	**b**	**c**	**d**	
Judge 1	Dimension A	0	2	4	6	
	Dimension B	6	2	4	0	*r* = –.80
Judge 2	Dimension A	6	2	4	0	
	Dimension B	0	2	4	6	*r* = –.80
Aggregated ratings	Dimension A	3	2	4	3	
	Dimension B	3	2	4	3	*r* (CA) = +1.00
						r (AC) = –.80

moral outrage, Nickerson argues that it leaves out whether moral outrage is highly correlated to WTP at the level of the individual. She also provides the formula relating aggregated correlations and correlations of aggregates (see Appendix).

A between–group averaging design such as Monin's (2003) Study 1 assumes two things: that participants agree on ratings of attractiveness and familiarity (the intraclass correlation coefficient for attractiveness was .94, and for familiarity it was .78.), and also that the correlation between the two dimensions is not stronger within individuals than it is across individuals. If the former assumption is violated, then the correlation of averages is an overestimation of the average of correlations. If the latter assumption is violated, then the correlation of averages is an underestimation of the average of correlations.

THE PRESENT STRATEGY

The problem with using exclusively aggregated data is that it does not take into account inter–individual differences and in effect treats sample aggregates as population values. This article endeavors to re–introduce variability in the estimate of the beau-

TABLE 2. Overview of Studies

	Attractiveness		Familiarity		Correlations	
	Source	*n*	Source	*n*	CA	AC
Study 1	Aggregate	34	Individual	40	.64	.20
	Individual	34	Aggregate	40	.64	.37
Study 2	Aggregate	> 2,500	Individual	37	.48	.14
Study 3	Individual	39	Individual	39	.52	.25

Note. CA = Correlation of Averages; AC = Average Correlation

tiful–is–familiar effect. We proceed in three steps laid out in Table 2. In Study 1, we re–analyze results from Monin (2003), alternating between individual ratings of familiarity and averaged ratings of attractiveness and vice–versa. In Study 2, we keep familiarity individual, but now for attractiveness we use an aggregate based on a large enough sample ($n > 2{,}500$) to be more confident that the aggregate value is close to the population value. In Study 3, we sidestep the issue of aggregation altogether by relying purely on individual measures and computing correlations within individuals.

STUDY 1: REANALYSIS OF MONIN (2003)

We started by re–analyzing the results of Monin's (2003) Study 1 from two different angles: first we kept attractiveness aggregated and let familiarity vary at the level of the individual, and then we kept familiarity aggregated and let attractiveness vary at the level of the individual. Our goal was to compare the averaged correlations using each of these two approaches to the already published correlation of averages. Given unavoidable idiosyncratic differences in people's ratings of attractiveness and familiarity, we predicted that the correlation of averages was an overestimate of the individual values (see Appendix). Thus, allowing one dimension to vary at the level of the individual and calculating the average of correlations should lead to lower correlation coefficients than Monin's (2003) original estimates. However, if the distribution of these coefficients is reliably greater than zero, it would still

provide support for the "beautiful-is-familiar effect" under more conservative conditions.

METHOD

Eighty pictures (40 from each gender) were taken from a yearbook and arranged on four sheets. Participants rated the 80 pictures on a 1 to 10 scale either on attractiveness ($n = 34$) or on familiarity, defined as the confidence that they had seen the person on the picture before ($n = 40$). Participants rating familiarity were led to believe that half the pictures were of students still on campus. In reality all pictures were taken from years prior to our respondents' presence on campus; thus the targets were all new.

RESULTS

Correlation of Averages. As in Monin (2003), we started by computing for each picture the average attractiveness across 34 judges and the average familiarity across 40 judges. When these 80 pairs are correlated using picture as the level of analysis we find, as previously reported, a correlation of averages that is quite high, $r = .64$, $p < .001$.

Average of Correlations. To move one step away from aggregates, we first took the average attractiveness for each picture, used that as a normed value, and generated the correlation between each of the 40 judges' familiarity ratings and this norm score. This produces 40 correlation coefficients. The average of these correlation coefficients is much lower than the correlation of averages, $M = .20$, $SD = 15$, though significantly greater than zero, 95% C.I. = [.16; .26]. Observing the distribution of correlation scores (see Figure 1a) reveals that whereas some participants exhibit high correlation coefficients that look like the correlation of averages, others show a much lower association between individual familiarity and aggregated attractiveness, with some respondents even exhibiting a negative correlation coefficient.

The other way to approach this data is to take the average familiarity for each picture, use that as a normed value, and generate the correlation between each of the 34 judges' attractiveness rat-

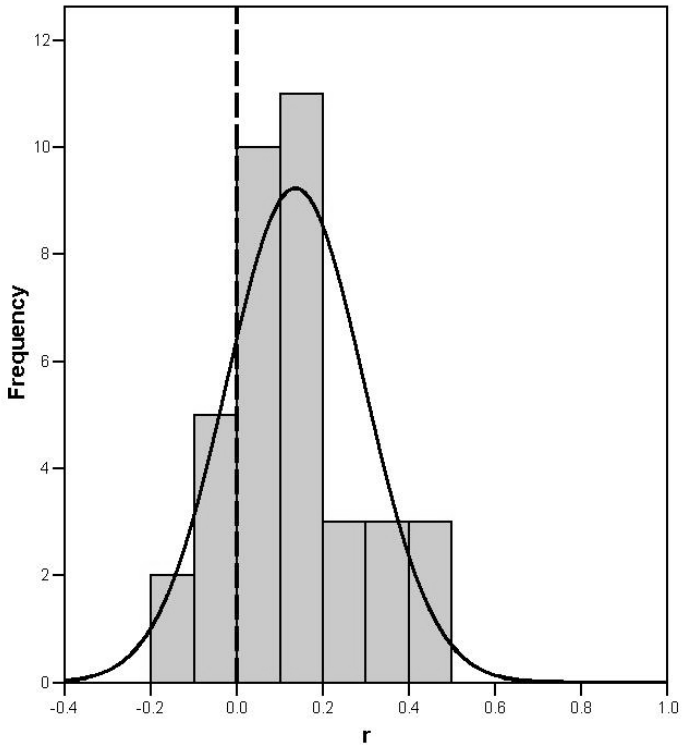

FIGURE 1a. Distribution of individual correlations between aggregated attractiveness (n = 34) and individual ratings of familiarity (Study 1).

ings and this norm score. This produces 34 (all positive) correlation scores (see Figure 1b), $M = .37$, $SD = .11$, with a 95% C.I. of [.33; .41].

DISCUSSION

This re–analysis provides an informative reconsideration of previous data. Such a mixed analysis seemed a salutary first approach to the issue of aggregation. We kept one dimension aggregated, but let the other vary with each participant. As discussed by Nickerson's (1995) analysis, and given some predictable disagreements in rat-

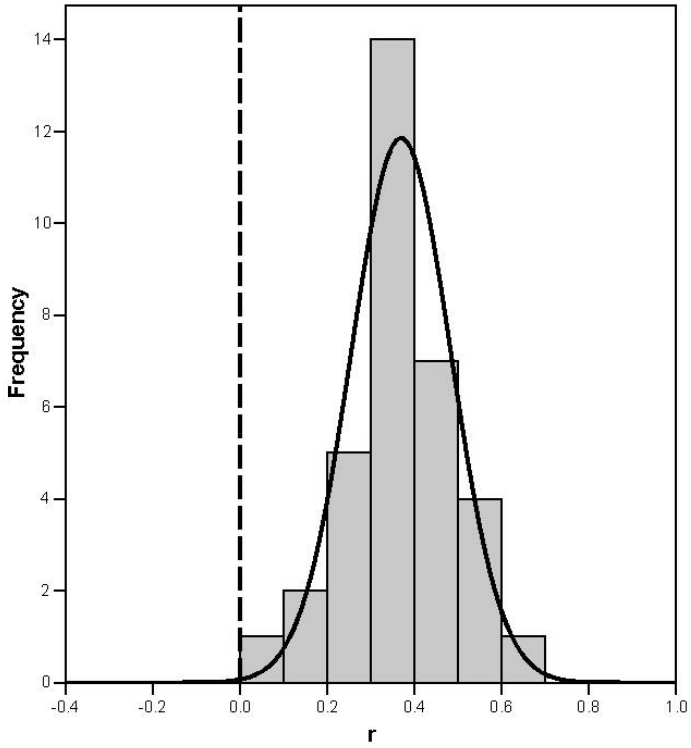

FIGURE 1b. Distribution of individual correlations between aggregated familiarity (n = 40) and individual ratings of attractiveness (Study 1).

ings among judges, average correlations were lower than correlated averages. This suggests that the correlation reported in Monin (2003) may be an overestimation of the link between attractiveness and familiarity at the individual level. However, the average correlation was significantly greater than zero. Thus we demonstrated the beautiful–is–familiar effect even with this more conservative test.

We used "mixed" correlations in this re–analysis, with one dimension aggregated and one at the level of the individual. The logic of aggregation is to get an estimate of a population norm. Although this first demonstration has the desirable feature that we can switch around which dimension was aggregated, it would be preferable in order to compute mixed correlations to obtain first

normed values from a larger sample to obtain a better estimate of population values. Aggregates in Study 1 are based on 40 respondents at the most. Study 2 again uses the logic of mixed correlations but uses attractiveness ratings from a much larger sample to address this issue.

STUDY 2: AM I HOT OR NOT?

This second study used a mixed design similar to that used in Study 1, but used a much larger sample to generate a norm of attractiveness ratings. The larger sample was gathered by relying on the website amihotornot.com, on which users post their photograph to be rated on a single scale of attractiveness by thousands of visitors. Unlike Study 1, this study does not enable us to alternate which dimension is aggregated in the analysis (because we do not have access to individual ratings of attractiveness), but the trade–off is the greater reliability of the attractiveness ratings given the great number of respondents.

METHOD

Participants. Thirty–seven Stanford students took part in this experiment for course credit. We removed one participant who did not seem to believe the cover story and answered "1" on all 90 pictures, yielding no usable variability.

Materials and Procedure. Participants engaged in a bogus recognition paradigm (similar to Monin, 2003, Study 4) employing photographs taken from the site amihotornot.com[2]. Ninety photographs that had each received over 2,500 votes were selected, with ten at each level of attractiveness ranging from 1.9 to 9.9 in one–unit increments. These photographs were then divided into two sets of 45 (with five pictures at each level of attractiveness), which were presented in counterbalanced order

2. Amihotornot.com [http://www.amihotornot.com] is a website launched in 2000 by James Hong and Jim Young on which visitors can view photos posted voluntarily by fellow users and rate their comeliness on a scale from 1 (Not) to 10 (Hot). On 6/28/04, the site claimed that it had received 12,200,000 photos and 8 billion votes. It has received numerous media mentions and has become a popular culture phenomenon.

to different groups of participants. Participants were told that they were in an experiment on subliminal perception, that they would first see photographs below the threshold of conscious perception, and then would be shown several photographs and would have to guess which they saw. In reality, they were only shown multicolored masks flashed on the screen in rapid succession. All photographs at test were new. Participants rated each photograph's familiarity on a scale from 1 (*not at all familiar*) to 10 (*very familiar*). It is worth noting that these naturalistic photographs were much richer in content than stimuli ordinarily used in face recognition, including not only differences in grooming and expression, but also in framing (some including the poser's body), and context.[3]

RESULTS

Correlation of Average. We observed the same type of correlations between familiarity and liking that we had encountered in previous studies. When we averaged ratings of familiarity across participants and correlated this average with the normed attrac-

3. This study included a between–subject manipulation of mood after the first 45 test trials: Twenty participants saw a five–minute clip of the television cartoon show *The Simpsons* (positive), while 17 participants saw a nature video (neutral) of approximately the same length. Although *The Simpsons* indeed increased participants' reported mood state, $F(1,35) = 6.2$, $p < .02$, it did not have much impact on familiarity ratings. We conducted a Mood × Order analysis of covariance on average familiarity ratings in the second block, with average familiarity ratings in the pre–movie recognition task as a covariate, and found no effect of mood, nor an interaction with order, both $F(1,31) < 1$. Note that the mood manipulation did not impact significantly correlations collected after the manipulation, $t(34) = 1.2$, *ns*.

Though it is not the focus of this article, it is worth discussing the absence of mood effects. In light of findings by Monahan, Murphy, and Zajonc (2000) showing that repeated exposure leads to generalized mood improvement, we expected that improving people's mood might increase their general sense of familiarity. Yet, despite its ability to change people's reported mood state, our mood manipulation did not lead to higher familiarity ratings: Improved mood was not misattributed to increased familiarity for stimuli subsequently presented. These results suggest that the warm glow heuristic may rely on a diffuse positive feeling attached to a given stimuli, but that it shows some specificity as to the source of the feeling. In other words, positive affect might be attributed to the wrong *feature* of the stimulus, but it may still need to come from the stimulus itself (for another example of the limits of misattribution, see Winkielman, Zajonc, & Schwarz, 1997). Given its tangential nature and to simplify our presentation, we ignore this variable in the rest of the article.

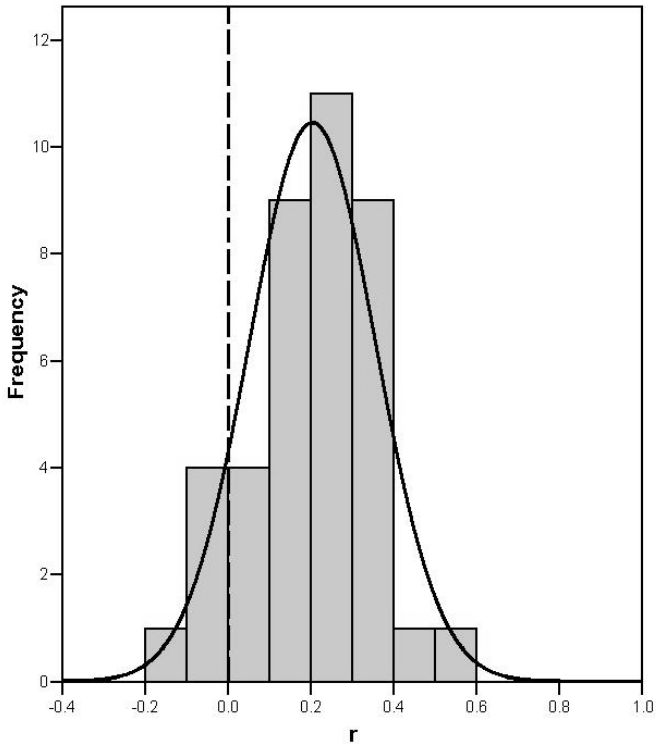

FIGURE 2. Distribution of individual correlations between aggregated attractiveness (n > 2,500) and individual ratings of familiarity (Study 2).

tiveness from the website, we observed a high positive correlation (r = .48, p < .001).

Average of Correlations. When we computed an individual correlation for each participant between their 90 individual ratings of familiarity and the normed attractiveness scores, we found individual correlations that ranged from –12 to +.49 and an average correlation that was much lower than the correlation of averages (M = .14, SD = .16, see Figure 2). This average correlation was again significantly different from zero, $t(35) = 5.2$, p < .001, and the 95% confidence interval for the population coefficient was [+.09; +.19].

DISCUSSION

Study 2 replicates the phenomenon documented in Monin (2003) using new stimuli, a new procedure, and ratings of attractiveness from over 2,500 respondents for each picture. As discussed by Nickerson (1995), the correlation of averages between familiarity and attractiveness (r = .48) in this study was quite different from the individual correlations (mean r = .14), defined as correlations between individual familiarity ratings and a norm of attractiveness established over a great number of respondents. Although these individual correlations were smaller, we were heartened to find out that they were, on average, still significantly greater than zero. These data provide further support for a beautiful–is–familiar effect, albeit a smaller one than was observed when the correlation of averages was the only statistic assessed.

STUDY 3: WITHIN–INDIVIDUAL CORRELATIONS

Study 2 presented individual correlations between individual ratings of familiarity and aggregated ratings of attractiveness; Study 3 proposes to take one additional step, correlating ratings of attractiveness and familiarity within individuals.

METHOD

Participants. Fifty–one Stanford students took part in a mass questionnaire session for course credit. Seven participants had to be excluded because they returned incomplete surveys. Another five participants did not express familiarity for any face in the familiarity task, literally giving a rating of 1 for each of the 60 stimuli. The lack of variance in these ratings precluded the computation of a correlation score so we excluded these five participants from our analyses. The subsequent analyses are based on the remaining 39 participants.

Materials and Procedure. We selected 60 yearbook faces randomly from the ones used in Monin (2003) and presented participants with essentially the same instructions as in Monin's (2003) Study 1, except that every participant first rated the familiarity of

each face in the set and then rated the attractiveness of every face in the set. Familiarity was always measured before attractiveness because after having rated attractiveness, all faces would have been seen and therefore would be familiar. Two versions were administered. In the *immediate* version, both ratings were presented back to back: The first three pages comprised the familiarity ratings and the following three pages comprised the attractiveness ratings. In the *delayed* version, the two blocks were separated by ten pages of unrelated questionnaires. For both the familiarity and attractiveness ratings, participants indicated their ratings in a box under each picture using a score from 1 to 10.

RESULTS

Correlation of Averages. We started by computing, within each condition, average scores of attractiveness and familiarity for each face and correlating these two matrices of 60 scores. This yields a correlation of averages of $r = .42$, $p < .001$, without delay, and $r = .53$, $p < .001$, with delay—overall $r = .52$, $p < .001$.

Average of Correlations. Because we collected ratings of familiarity and attractiveness within participants, we were able to compute correlation scores for each participant and to treat those as individual pieces of data. Without delay, the average correlation was .27, with a 95% confidence interval of [.12; .42], and with delay it was .23, 95%, C.I. = [.11; .36]. The average correlation was not different between the two conditions, $t(37) = -.33$, *ns*. Overall, the average correlation was $r = .25$, and the 95% C.I. over the 39 participants was [.16; .34] (Figure 3).

DISCUSSION

In contrast to prior studies, this experiment collected both familiarity and attractiveness ratings made by the same raters about the same stimuli, enabling us to address concerns about the inferences drawn from correlations of averages. Although the average correlation ($r = .25$) was lower than the correlation of averages ($r = .52$), the former was still significantly higher than zero. Therefore, even computed purely within individuals, the

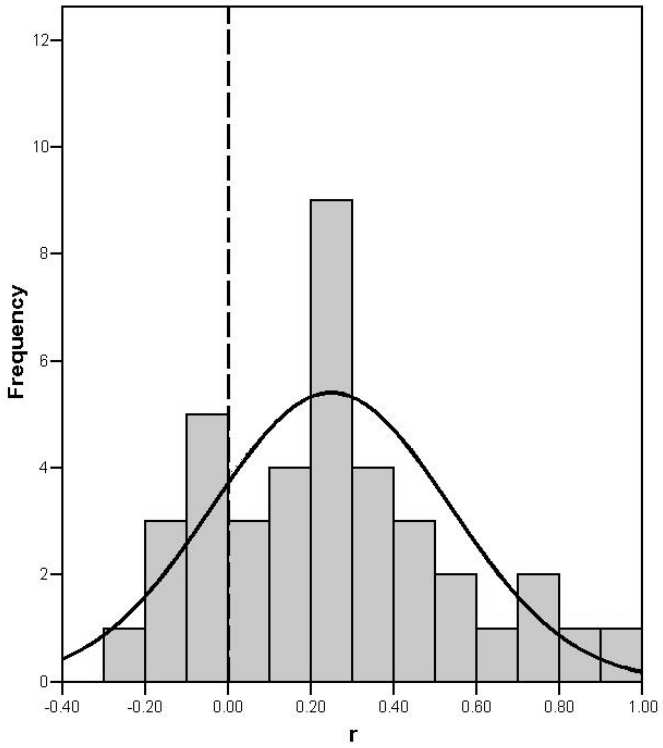

FIGURE 3. Distribution of within-subject correlations between individual attractiveness ratings and individual familiarity ratings (Study 3).

correlation between attractiveness and familiarity was on average positive, and it did not matter whether the attractiveness ratings were collected immediately following the familiarity ratings or ten pages later.

GENERAL DISCUSSION

Our goal in this paper was really twofold: On the one hand, we wanted to address a potential weakness in Monin's (2003) demonstration of the warm glow heuristic, which relied mostly on aggregate ratings. On the other hand, we wanted to illustrate the difference between a correlation of averages and an average of

correlations, a simple distinction that does not have the currency one would expect in experimental psychology (Nickerson, 1995). The studies reported above achieve both goals.

Study 1 re–analyzes the data from Monin's first study, alternating whether familiarity or attractiveness was treated individually while the other dimension was aggregated. Study 2 used attractiveness ratings from a large sample of respondents on a public website as a norm, which were then correlated with individual ratings of familiarity. In Study 3 we collected both attractiveness and familiarity ratings from each individual, enabling us to correlate individual ratings of attractiveness with individual ratings of familiarity. While in Study 3 ratings of attractiveness invariably came after ratings of familiarity, in the first two studies these was no possibility that one rating would contaminate the other.

Together, these three studies show that our methodological concerns were well–founded. As is shown in Table 2, correlations of averages in the studies ranged from .48 to .64, whereas averages of correlations ranged from .14 to .37. In all cases, the former were higher than the latter. In all cases, however, the average correlation was significantly greater than zero. Thus these data still clearly support the prediction that attractive faces look more familiar. The more attractive a face, the more familiar it seemed. These new findings strengthen a growing body of evidence illustrating the beautiful–is–familiar effect and, more generally, reflecting the operation of the warm glow heuristic (Monin, 2003; Corneille et al., in press).

UNDERSTANDING THE WARM GLOW HEURISTIC

Often, calling something a heuristic describes a relationship between two variables more than it posits a mechanism underlying this relationship. After all, the word "heuristic" has been used to describe a number of very distinct cognitive processes (Gigerenzer, Todd, & the ABC Research Group, 1999). It is therefore important to go beyond naming the heuristic and to specify the mechanism responsible for the effect. We now discuss possible mechanisms underlying the beautiful–is–familiar effect.

Kahneman and Frederick (2002) proposed that a general mechanism underlying heuristics was "attribute substitution," a meta–cognitive process in which a hard question (assessing an attribute that is hard to assess, such as frequency) is answered as if it were another, easier one (e.g., assessing representativeness instead). Indeed, the warm glow heuristic seems to come into play most when assessing familiarity becomes hard (Monin, 2003, Study 5). When there was no delay and participants anticipated testing, the effect was only apparent among new faces (false alarms). When the recognition task was unexpected and came after a 25–minute delay, it was much harder to recognize any face, and the impact of attractiveness was observable among both new faces (false alarms) and old ones (hits). In line with attribute substitution, when participants could not rely on clear memory traces, they turned instead to the next best thing, and one attribute that is always quickly assessed is how much one likes the stimulus (Zajonc, 1980, 1998).

Kahneman and Frederick's (2002) attribute substitution model fits our theorizing about the warm glow heuristic. Lacking a strong memory trace, and maybe the cognitive resources and motivation to make a direct recognition judgment, participants tapped into their liking for stimuli to infer familiarity, and in some studies, to guess whether they had seen the stimulus before. This model assumes that participants make an implicit connection between positive affect and familiarity. One possible origin of this connection is the fact that familiar things are indeed liked more (Zajonc, 1968), so an implicit understanding of the mere exposure effect would justify using liking as a proxy for prior exposure. Another possibility is that both dimensions can be reduced to the common denominator of perceptual fluency (Jacoby & Dallas, 1981): On the one hand, the ease with which one processes a stimulus is used as an indicator of prior exposure or familiarity (Johnston, Dark, & Jacoby, 1985), especially when this ease is unexpected (Whittlesea & Williams, 2001). On the other hand, this same fluency is a strong predictor of liking and judgments of beauty (Reber, Winkielman, & Schwarz, 1998; Schwarz, 2004; Reber, Schwarz, & Winkielman, 2004). Although this interpretation accounts for only part of the existing data (in particular it cannot explain the good–is–familiar effect with positive words—Monin, 2003, Study 4), it may explain the implicit

association between liking and familiarity that seems to underlie our participants' judgments.

An alternative to the attribute substitution account (Kahneman & Frederick, 2002) would be a weighted additive process (e.g., Beattie & Baron, 1991; Keeny & Raifa, 1976). In this account, liking is one of many cues for familiarity, and those cues are combined to lead to an overall judgment. If other cues are absent (e.g., after a time delay, see Monin, 2003, Study 5) people will give more weight to the remaining cues and will weight cues such as affect and fluency more heavily. The present data may be better accounted for under a weighted additive model than with a simpler version of attribute substitution. While the data clearly show that likeability has an influence on judgment, the correlations (especially with the correctives applied in this article) are fairly low. When attributes are substituted, one would expect much higher correlations (Kahneman & Frederick, 2002). If people are simply substituting likeability for familiarity, within–subject judgments of the two attributes should be nearly identical.

Yet another possibility would be that the beautiful–is–familiar effect is simply another example of a more general halo or beautiful–is–good effect (Dion et al., 1972). According to this view, familiarity is just one of the many positive features, such as intelligence, success, social skills, or maturity, that attractive people are assumed to possess. One version of this interpretation predicts that we should obtain the same type of correlation between attractiveness and any positive personality trait as we do with familiarity. To test this possibility, Monin (2003, Study 1) asked a third group of participants ($n = 36$) to rate the maturity of the pictures. The correlation of averages between attractiveness and maturity ($r = .29$, $p < .01$) was significantly lower than with familiarity ($r = .64$, $p < .01$), Fisher's $z = 2.85$, $p < .01$, familiarity and maturity did not correlate ($r = .00$), and partialing out maturity did not reduce the correlation between attractiveness and familiarity ($r = .66$, $p < .01$). Thus even though the link between attractiveness and maturity reflects a possible halo effect, that effect seemed quite independent from the link between attractiveness and familiarity.

Another version of this halo interpretation contends that the effect results from evaluative matching between the stimulus and the

response. If the response to be produced ("familiar") is positive in valence, it may be easier to produce it after being exposed to a positive stimulus. This behavioral facilitation could be argued to be an artifact in the method used by Monin (2003). To test this possibility, Corneille et al. (2004) replicated Monin's Study 2 while manipulating whether participants indicated having seen a face before by choosing a positive (congruent) image or a negative (incongruent) one. The effect did not disappear in the incongruent condition (indeed it seemed stronger), ruling out the evaluative matching interpretation. Note also that although familiarity, loosely defined, is likely to be semantically associated with positivity, in the present studies it was explicitly defined to participants as confidence that one had seen a face before on campus (Studies 1 & 3) or earlier in the experiment (Study 2). Similarly, it is less credible that the word "old" that was used to indicate prior exposure in the recognition studies (e.g., Monin's [2003] Studies 2, 4, & 5) possesses an inherently positive quality that would make it easy to explain the effects away. All of these data, taken together, suggest that the beautiful–is–familiar effect is unlikely to be merely the result a of halo effect by which familiarity is taken as just another way to measure goodness. Instead, we propose that it is the result of a meta–cognitive shortcut (the warm glow heuristic) whereby liking is taken as a cue for familiarity.

THE WARM GLOW HEURISTIC BEYOND AGGREGATION

In retrospect, the high correlation scores obtained in prior research (e.g., $r = .64$, in Study 1) appear less reflective of the effect size of the association between attractiveness and familiarity at the level of the individual than the individual correlations computed in this article. Furthermore, this approach opens new roads for empirical inquiry. It is probable that the variability observed in individual correlations does not reflect solely error, but that individuals differ in meaningful ways in the extent to which they rely on the warm glow heuristic when making familiarity judgments. Future research should identify which individual differences predict greater susceptibility to beauty in the assessment of familiarity.

In addition, a growing literature demonstrates how people's use of affective and fluency–based heuristics is influenced by causal reasoning (Schwarz & Clore, 1983; Oppenheimer, 2004). For example, Oppenheimer (2004) asked people to make judgments about surname frequency—a domain in which people typically use the availability heuristic (Tversky & Kahneman, 1973)—and showed that people did not use availability when the names were famous. When there was an obvious cause for the meta–cognitive state of availability (in this case fame), people discounted availability as a cue in judgment. It seems plausible that similar causal reasoning may influence the use of the warm glow heuristic, either because of a question about attractiveness before the familiarity question (Hilton, 1990) or because stimuli conspicuously proclaim their attractiveness (e.g., supermodels) so that positive affect is correctly attributed to its rightful source. Future research should investigate this possibility.

At the methodological level, we hope that this article presents a useful demonstration for students and colleagues of the fallacy of equating correlated averages and averaged correlations, and illustrates some of the strategies available to address that issue. We want to emphasize once more that from a theoretical point of view both statistics are valid; our point is that investigators need to be fully aware of how their choice of a statistic affects which claims they are in a position to make. Correlated averages are more likely to make a point about features of stimuli and how they relate; average correlations are better estimates of effect sizes at the level of the individual. Whether the focus is on stimuli or on respondents determines which statistic should hold center stage. We think it useful, however, as we have in this project, to cast light on both perspectives. We hope that this article will inspire others do so with clarity and confidence.

APPENDIX

Nickerson (1995) provides a formula for the relationship between a correlation of averages and the average of within-subject correlations (we assume standardized scores here for simplicity):

$$r_{\overline{xy}} = \frac{(1/N)\overline{r_{xy}} + ((N-1)/N)\overline{r'_{xy}}}{\sqrt{(1/N) + ((N-1)/N)\overline{r'_{xx}}}\sqrt{(1/N) + ((N-1)/N)\overline{r'_{yy}}}}$$

N number of respondents

$\overline{r_{xy}}$ mean of within-respondent correlations

$\overline{r'_{xy}}$ mean of correlations between every respondent's x score and every other respondent's y score

$\overline{r'_{xx}}$ mean of correlations between every respondent's x score and every other respondent's x score

$\overline{r'_{yy}}$ mean of correlations between every respondent's y score and every other respondent's y score

This formula reveals that the average correlation and the correlation of averages are identical if r_{xy} equals r'_{xy} (i.e., if you get the same average correlation by randomly pairing any two observations or looking within-respondent) *and* both r'_{xx} and r'_{yy} equal 1 (i.e., judges are in perfect agreement on both dimensions). It also shows that the cross-respondent correlations carry much more weight (N-1 times more) in the correlation of average than does the within-respondent correlations. Finally, and most important, if *any single* of the r'_{xx} or r'_{yy} correlations falls below 1 (which you would expect with error alone), then the denominator will fall below 1, making it likely that a correlation of averages will be higher than an average of correlations.

REFERENCES

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluation from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*, 431–441.

Beattie, J., & Baron, J. (1991). Investigating the effect of stimulus range on attribute weight. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 571–585.

Chaiken, S. (1979). Communicator physical attractiveness and persuasion. *Journal of Personality and Social Psychology, 37*, 1387–1397.

Corneille, O., Monin, B., & Pleyers, G. (in press). Is positivity a cue or a response option? Warm–glow versus evaluative–matching in the familiarity for attractive and not–so–attractive faces. In press, *Journal of Experimental Social Psychology*.

Dion, K. K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24,* 285–290.

Downs, A. C., & Lyons, P. M. (1991). Natural observations of the links between attractiveness and initial legal judgments. *Personality and Social Psychology Bulletin, 17*, 541–547.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta–analytic review of research on the physical attractiveness stereotype. *Psychology Bulletin, 110*, 107–128.

Garcia-Marques, T., Mackie, D. M., Claypool, H. M., & Garcia-Marques, L. Positivity Can Cue Familiarity. *Personality and Social Psychology Bulletin, 30*, 585–593.

Gigerenzer, G., Todd, P.M., & the ABC Research Group (1999). *Simple Heuristics that Make us Smart.* New York: Oxford University Press.

Gordon, K. H. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology, 3*, 398–400.

Hamermesh, D. S., & Biddle, J. E. (1994). Beauty and the labor market. *American Economic Review, 84*, 1174–1195.

Hilton, D. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107*, 65–81.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*, 306–340.

Johnston, W. A., Dark, V. J., & Jacoby, L. L. (1985). Perceptual fluency and recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 3–11.

Kahneman, D., & Knetsch, J. L. (1992). Valuing public goods: The purchase of moral satisfaction. *Journal of Environmental Economics and Management, 22*, 57–70.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Keeney, R. L., & Raifa, H. (1976). *Decisions with multiple objectives: Preferences and value trade offs.* New York: Wiley.

Monahan, J. L., Murphy, S. T., & Zajonc, R. B. (2000). Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science, 11*, 462–473.

Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology, 85*, 1035–1048.

Nickerson, C. A. E. (1995). Does willingness to pay reflect the purchase of moral satisfaction? A reconsideration of Kahneman and Knetsch. *Journal of Environmental Economics and Management, 28*, 126–133.

Oppenheimer, D. M. (2004). Spontaneous discounting of availability in frequency judgment tasks. *Psychological Science, 15*, 100–105.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aes-

thetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8*, 364–382.
Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science, 9*, 45–48.
Rosenthal, R. (1987). *Judgment studies: Design, analysis, and meta–analysis*. NY: Cambridge University Press.
Schwarz, N. (2004). Meta–cognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14*, 332–348.
Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*, 513–523.
Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207–232.
Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory and Cognition, 20*, 291–302.
Whittlesea, B. W., & Williams, L. (2001). The discrepancy–attribution hypothesis: I. The heuristic basis of feelings and familiarity. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*, 3–13.
Winkielman, P., Zajonc, R. B., & Schwarz, N. (1997). Subliminal affective priming resists attributional interventions. *Cognition and Emotion, 11*, 433–465.
Zajonc, R. B. (1962). A note on group judgements and group size. *Human Relations, 15*, 177–180.
Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*, 1–27.
Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*, 151–175.
Zajonc, R.B. (1998). Emotions. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds), *The handbook of social psychology* (pp. 591–632). New York: McGraw–Hill.