In Press – Emotion. Please direct any concerns or errors to me. Thanks!

On the Association between Loneliness and Bathing Habits: Nine Replications of Bargh and Shalev (2012) Study 1

M. Brent Donnellan, Richard E. Lucas, Joseph Cesario

Michigan State University

Draft Date: 21 January 2014

Address correspondence to:

M. Brent Donnellan

Department of Psychology

316 Physics, Rm 252A

Michigan State University

East Lansing, MI 48824

donnel59@msu.edu

Thanks to R. Chris Fraley for helpful comments on an earlier draft.

Abstract

Bargh and Shalev (2012) hypothesized that people use warm showers and baths to compensate for a lack of social warmth. As support for this idea, they reported results from two studies that found an association between trait loneliness and bathing habits. Given the potential practical and theoretical importance of this association, we conducted nine additional studies on this topic. Using our own bathing/showering measures and the most current version of the UCLA Loneliness scale (Russell, 1996), we found no evidence for an association between trait loneliness and a composite index of showering/bathing habits in a combined sample of 1,153 participants from four studies. Likewise, the aggregated effect size estimate was not statistically significant using the same measures as the original studies in a combined sample of 1,920 participants from five studies. A local meta-analysis including the original studies yielded an effect size estimate for the composite that included zero in the 95% confidence interval. The current results therefore cast doubt on the idea of a strong connection between trait loneliness and personal bathing habits related to warmth.

Word Count: 177

Key words: Loneliness, Effect Sizes, Statistical Precision, Replication, Meta-Analysis

Bargh and Shalev (2012) hypothesized that lonely individuals use physical warmth as a substitute for a lack of social connection. They suggested that individuals "self-regulate their feelings of social warmth (connectedness to others) with applications of physical warmth (through taking warm baths or showers)" (p. 155). They conducted two correlational studies on this topic and reported substantial associations between loneliness and a "physical warmth" composite variable based on the average frequency, duration, and temperature of showers and baths (r = .57 and r = .37). They noted that this was a "strong" association in Study 1a and emphasized the amount of variance that loneliness explained when considering the physical warmth extraction variable.

> This index variable correlated with UCLA Loneliness Scale scores (r = .57, p <.0001) such that 32.5% of the variation in physical warmth extraction during bathing was explained by how lonely the participant was. Overall, in this student sample, chronic levels of "social coldness" were strongly related to the amount of physical warmth the individual consumed each week in the form of bathing." (Bargh & Shalev, 2012, p. 156).

This package of studies received popular attention including coverage in mainstream media outlets. Some of the discussion in the popular press focused on the potential palliative effects of warm water for feelings of loneliness. To be sure, Bargh and Shalev suggested that a connection between physical warmth and psychological attributes might serve as "a boon to the therapeutic treatment of syndromes that are mainly disorders of emotion regulation" (p. 155).

In light of the potential practical significance surrounding the association between bathing practices and loneliness, we attempted nine replications using sample sizes that were each substantially larger than the original report. We undertook this effort for at least three reasons. First and foremost, the original studies were based on small samples (n = 51 and n = 41) and therefore they generated effect size estimates with relatively wide confidence intervals

(see Cumming, 2012; in press). Specifically, the 95% confidence interval for the .57 correlation was approximately .35 to .73 whereas the 95% confidence interval for the .37 correlation was approximately .07 to .61. In other words, the range of plausible values from the original studies was large (see Bonett, 2012). Schönbrodt and Perugini (2013) suggested that effect sizes are fairly unstable in small samples and concluded that "there are few occasions in which it may be justifiable to go below n = 150" (p. 611). The precise magnitude of any effect size estimates matters for "real world" applications (Schimmack, 2012) and it is important to test whether outside researchers can obtain similarly sized correlation coefficients. Indeed, large effect sizes are relatively rare in social psychology (Richard, Bond, and Strokes-Zoota, 2003) and correlations around .50 are larger than many substantive correlations reported in the literature (see Hempill, 2003).

The second reason is related to the strength of the evidence in the original publication. The survey question asking about preferred water temperature is the one item directly related to theorizing about warmth and loneliness. Items about the frequency and duration of showers/baths refer to different kinds of behaviors and are not as relevant to the substitutability hypothesis with respect to physical warmth and psychological warmth. The correlation for the warmth item was not statistically detectable in Study 1a using a two-tailed test (r = .26, 95% CI = -.02 to .50) whereas it was statistically detectable in Study 1b (r = .34, 95% CI = .04 to .59). Thus, the strength of the evidence for the connection between water temperature and loneliness in the original report could be considered inconsistent by some standards.

The final reason concerns the raw data from at least one of the studies in the original report. Most notably, the distribution of the frequency of showering/bathing variable in Study 1a

was skewed. Upon inspection, we learned that 46 of the 51 participants (90%) reported taking less than one shower or bath per week. This level of response seemed unusual and indicates that the correlation with the loneliness scale was driven by those few participants who did not provide such an unusual response. In addition, the modal response to the temperature variable in Study 1a was also surprising given that 24 out of 51 participants selected the "cold" response (47%) and 18 selected the “lukewarm” response (35%). These potentially unusual responses suggest that results from Study 1a are unlikely to generalize to other samples.

In sum, there is a need for more research on the connection between personal bathing habits and loneliness. Accordingly, we conducted nine studies on this association. Before describing our attempts to duplicate the original results, we acknowledge some of the controversies associated with replication studies (see Asendorpf et al., 2013; Valentine et al., 2011). One issue concerns what it means to successfully replicate an initial finding (see Asendorpf et al., 2013, p. 112; Simonsohn, 2013). One standard for replication could be whether an effect is in the same direction and statistically significant in an additional sample. This criterion is potentially problematic because these kinds of judgments are tied to sample size considerations (i.e., the subsequent *p*-value is determined, in part, by sample size). Researchers can increase the odds of failing to detect an effect by selecting a sample size that was smaller than the original study. Conversely, researchers can increase the odds of finding virtually any correlation or mean difference statistically significant using a large enough sample size.

An alternative approach for judging the success of a replication study is to consider whether subsequent studies yield similar effect size estimates as the original. This approach has the advantage of orienting decisions around effect sizes rather than statistical significance levels

(see Cumming, in press). Nonetheless, there are challenges involved in making judgments of similarity (see Asendorpf et al., 2012; Simonsohn, 2013). Are correlations of .37 and .57 different? The question could be addressed statistically but whether differences in estimates are statistically significant depends, in part, on the sample sizes of the respective studies.

One possible solution to the issue of evaluating replication attempts has recently been proposed by Simonsohn (2013). The gist of his proposal is to consider failures to replicate to occur when the effect size estimate from a new study was too small to be detected using the original sample size. This perspective generates a rule of thumb recommendation that replication studies need 2.5 times as many participants as the original. The advantage of Simonsohn's 2.5 times proposal is that it motivates replicators to use larger sample sizes, a suggestion that will provide the literature with more precise parameter estimates.

In addition to the controversial issue of making judgments about the outcome of a replication attempt, there is discussion as to how to classify replication studies. Asendorpf et al. (2013) define replicability as a situation in which "a finding can be obtained with other random samples drawn from a multidimensional space that captures the most important facets of the research design" (p. 109). As applied to current correlational study, a replication should therefore include similar participants and measures as the original study. As before, however, the issue of similarity entails a number of judgment calls about the nature of the procedures, measures, and participants.

In light of these considerations, we make the following stipulations. First, we note when the effects are statistically detectable in each of our studies with the caveat that each of our studies had more statistical power than the original Bargh and Shalev studies (each met the 2.5

times the original sample size rule of thumb proposed by Simonsohn, 2013). We also consider whether we repeatedly get similar effect sizes as the Bargh and Shalev (2012) report. We evaluate this issue in each study individually and then by aggregating across our multiple studies. Second, we make a distinction between near-exact and exact replications on the basis of measurement. When we use the exact same items, we consider our studies exact replication attempts. When we use slightly different items, we consider our studies near exact replication attempts. This will allow us to evaluate whether there is any substantial difference in results between studies with different measures.

We should emphasize that the original authors do not directly address issues of sampling and measurement as moderators of the underlying effects. Bargh and Shalev (2012) used convenience samples of college students and members of a New England community and reported that these two studies support the claim that "people tend to substitute physical warmth for the social warmth that is missing from their lives" (p. 160). Thus, we should be able to obtain results in multiple kinds of convenience samples. In terms of their measures of bathing habits, Bargh and Shalev created their own items and used a loneliness scale modeled on the measure developed by Russell (1996). There is no theoretical reason why slight changes in item wording or response options should make it difficult duplicate the relevant effects. Moreover, if relevant effects are highly contingent upon minor issues in wording, it would weaken claims about real-world or potential therapeutic implications of the results. To be sure, the authors of the original study proposed no theory-based limitations to the generality of the reported effects.[1]

---

[1]Difficulties duplicating the original finding may mean that there was something distinct about the original methodology that led to valid results that do not generalize to slightly different contexts. These kinds of results are

Overview of our Studies

Studies 1 to 4 are near-exact replications whereas Studies 5-9 are exact replications of the loneliness and bathing habit studies. These studies are presented in chronological order. Studies 1 to 4 use slightly different measures than those used by Bargh and Shalev (2012). For example, the response options for our showering and bathing items differed slightly from those used by Bargh and Shalev (both sets of item responses are reported in the appendix). In addition, we used the complete 20-item UCLA Loneliness Scale version 3 whereas Bargh and Shalev (2012) used a 10-item modified version of the first version of the UCLA scale (Russell, Peplau, & Ferguson, 1978). Russell (1996) revised the original UCLA scale to include both negative and positively keyed items to address concerns about response biases (i.e., all of the items were keyed in the same direction; see Russell, 1996, p. 21) and to modify the wording of particular items. We had difficulty duplicating the original effect sizes in Studies 1 to 4 so we switched to the same measures used by Bargh and Shalev (2012) for studies 5 to 9. Each of our studies had over .98 power to detect a correlation of .30 or higher at $p < .05$ using a two-tailed test. To make binary decisions regarding statistical significance we used the standard $p < .05$ (two-tailed) convention with the caveat that we think effect sizes are more important than levels of statistical significance (see also Cumming, in press).

Study 1: College Students

Sample and Procedure

---

useful if proponents of a specific theory wish to identify boundary conditions of their original phenomena. Repeated difficulties in duplicating the original finding might therefore spur the original authors to research theoretically-based moderators of the original effect.

Participants were 235 college students (68.5% women; M age = 19.66 years, SD = 2.18) who received course credit as part of the Michigan State University Psychology Subject Pool during the Spring Semester of 2012. The majority of participants self-identified as white (84.2%). Data were collected as part of a larger study investigating associations between personality and expressive writing. Participants came to the lab and were greeted by trained research assistants. Participants then provided informed consent and all survey items were administered via computers using a web-based survey interface. They were randomly assigned to complete either the writing tasks first (n = 122) or the survey items first (n = 113). There were no differences between the two groups in terms of the primary results (see below) or mean levels of the variables in question (minimum p = .51). Survey items were administered in a fixed format such that all participants completed all survey items in the same order.

The measures analyzed here were a subset of the individual difference measures completed by participants including a measure of the Big Five domains. Embedded within the Big Five personality questionnaire were two "instructed response" items designed to identify careless responders (see Meade & Craig, 2012). Careless, inattentive, or random responding may impact effect size estimates from survey research (Credé, 2010; Maniaci & Rogge, 2014; Meade & Craig, 2012). One item asked participants to select the "moderately inaccurate" response option whereas a second item asked participants to select the "moderately accurate" option. Only participants who answered correctly to both items were included in the current analyses (approximately 76% of the participants in the larger writing study).

Measures

Loneliness was measured using the 20-item UCLA Loneliness scale (Version 3; Russell, 1996). Participants responded using a 4-point Likert-type scale (1= Never to 4 = Always). The reversed scored items were coded in the direction of higher loneliness prior to computing the scale composite scores (M = 1.99; SD = .54, Alpha = .94). *Showering/Bathing Behaviors* were assessed using item stems and responses designed to closely match the published descriptions in Bargh and Shalev. These items are listed with their distributions in the appendix. The three items were standardized to compute the composite Physical Warmth Index following the same procedures used by Bargh and Shalev (alpha = .01).

Results

There was no statistical association between Loneliness and the Physical Warmth Index (r = -.06, p = .348, n = 235; 95% CI = -.19 to .07) when using respondents who passed the quality-control checks. The same result was obtained when we used all participants with relevant data (r = -.04, p = .442, n = 310, 95% CI = -.15 to .07). The ordering of the tasks did not moderate the reported associations (Writing First: r = -.05, p = .584, n = 122; Survey First: r = -.08, p = .430, n = 113; z for difference = 0.19). Separate analyses with each shower/bath item did not support the original predictions (see Table 1). If anything, lonelier students are likely to report taking fewer showers. The hypothesis relevant correlation between the water temperature item and the loneliness scale was not statistically distinguishable from zero (r = -.025, p = .700, n = 235, 95% CI = -.15 to .10).

Discussion

Study 1 was a near-exact replication of Study 1a reported in Bargh and Shalev (2012) using a sample size that was more than four times larger than the original (n = 51 college students; 26 women; M age = 20.11 years, SD = 4.17). In contrast to the original results, we found a trivial correlation in the opposite direction of the correlation reported in Bargh and Shalev. Thus, we did not duplicate their original result with college students using either the significance test criterion or the effect size criterion. One difference between this study and the original concerns the specific measures used in the research as we noted in the Introduction. A second issue is that they randomly presented participants with either the lifestyle habits survey or the loneliness questions first whereas we used a fixed order for the survey questions. There was no indication in Bargh and Shalev that the kind of ordering impacted results but we investigated this possibility in Study 2 using a non-subject pool sample.

Study 2: Participants from Amazon.com's Mechanical Turk

Sample and Procedure

Participants were "workers" recruited through Amazon.com's Mechanical Turk (mTurk see Behrend, Sharek, Meade, & Wiebe, 2011; Buhrmester, Kwang, & Gosling, 2011; Mason & Suri. 2012) who responded to a task called Social Behaviors. Recruitment occurred between August 21, 2012 and August 24, 2012. This sample was intended to approximate the community sample Study 1b in Bargh and Shalev (n = 41; 16 women; M age = 2.60 years, SD = 11.49) who were recruited from the town green of a city in New England. Existing research suggests that mTurk samples provide data that are as reliable as data from college student subject pools while also being more diverse, at least in terms of age (Behrend et al., 2011; Buhrmester, et al., 2011). Participants were compensated 50 cents for responding to the survey and

participants had the option of skipping any question that they wished without loss of payment. We randomly assigned participants to complete either the bathing/showering items first or the loneliness questions first. Participants then completed demographic items and rated a series of adjectives for an unrelated study. The entire task took approximately 10 to 15 minutes. The web-based survey interface for Study 2 was the same as Study 1.

We included two directed response items to detect careless responding. One item embedded in the lifestyle questionnaire asked participants to select the "rarely" response option whereas the penultimate item embedded in the loneliness questionnaire asked participants to select "sometimes". We also included an item at the end of the demographic questions: "I responded to this survey honestly." Only the 480 participants who answered "Yes" to the honesty question and responded correctly to both careless responding items were included in the current analyses (59.6% women; M age = 33.18 years, SD = 12.39). This represents approximately 92% of the data collected from mTurk for this study. As with Study 1, participants largely self-identified as White (78.9%). Measures were the same as Study 1**.** The physical warmth index was created by standardizing the three bathing/showering items and averaging them into a composite (alpha = .13).

Results and Discussion

There was no evidence for a statistically detectable association between Loneliness (M = 2.28; SD = .58, alpha = .94) and the Physical Warmth Index (r = -.01, p = .902, n = 480; 95% CI = -.10 to .08). The same pattern was obtained when we used all participants with relevant data (r = -.06, p = .17, n = 522, 95% CI = -.15 to .03). The ordering of survey questions did not moderate the reported associations (Lifestyle First: r = -.05, p = .424, n = 254; Loneliness First: r

= .04, p = .505, n = 226; z for difference = 1.03) or generate mean-level differences in the two variables (maximum d = |.07|, minimum p = .46). Separate analyses with each shower/bath item also did not support the substitutability prediction (see Table 1). If anything, lonelier mTurk responders are likely to report taking fewer showers. The hypothesis relevant correlation between the water temperature item and the loneliness scale was not statistically distinguishable from zero (r = .02, p = .718, n = 479, 95% CI = -.07 to .11). In short, we duplicated the null result from Study 1 in a second sample drawn from a different population using a sample size that was almost 12 times as large as the sample size from Bargh and Shalev (2012) Study 1b. Nonetheless, we made a third attempt using a different procedure for obtaining participants.

Study 3: Participants from SurveyMonkey Audience

Sample and Procedure

Participants were recruited through SurveyMonkey audience, a panel maintained by an internet market research company (see http://help.surveymonkey.com/app/answers/detail/a_id/6666). Participants in Study 3 completed the survey in exchange for donations to charitable causes, entries into drawings for prizes, and points used for merchandise. Recruitment occurred between September 7, 2012 and September 12, 2012. Measures were the same as Study 1 and Study 2**.** We counterbalanced the order of the questions so that participants were randomly assigned to either complete the bathing/showering items first or the loneliness questions first. Participants then completed demographic items.

We contracted for 205 participants (four times the largest sample size in Bargh and Shalev) and SurveyMonkey collected 224 complete responses. SurveyMonkey calculated the

response rate at 28.6% based on the ratio of the number of completed surveys to the total number of invites sent to their panelists. The average time to complete the survey was 4 minutes and 37 seconds. As in Study 2, we only included the participants who answered "Yes" to the honesty question and correctly responded to both careless responding items (n = 210, 47.6% women; M age 47.17 years, SD = 15.98). This represents approximately 94% of the completed surveys. As with the other studies, participants largely self-identified as White (87.5%). Measures were the same as Studies 1 and 2 and the Physical Warmth index was created by standardizing the three bathing/showering items and averaging them into a composite (alpha = .17).

Results and Discussion

There was no evidence for a statistical association between Loneliness (M = 2.07; SD = .48, alpha = .92) and the Physical Warmth Index (r = .13, p = .063, n = 210; 95% CI = -.01 to .26) using our alpha criterion. The same pattern was obtained when we used all participants with relevant data (r = .10, p = .130, n = 224, 95% CI = -.03 to .23). Separate analyses with each shower/bath suggested that lonelier participants took longer showers but not warmer or more frequent showers (see Table 1). The hypothesis relevant correlation between the water temperature item and the loneliness scale was not statistically distinguishable from zero (r = -.015, p = .830, n = 210, 95% CI = -.15 to .12). One concern with Study 3 is that the p value for the index was .063 and the correlation was in the direction that would support the Bargh and Shalev hypothesis. Some readers may argue that we were close to replicating their result for the composite variable. Accordingly, we conducted an additional study to evaluate whether this effect would attain statistical significance at the conventional .05 level in a second sample of participants drawn from the SurveyMonkey panel.

Study 4: Participants from SurveyMonkey Audience

Sample and Procedure

As in Study 3, participants were recruited through SurveyMonkey audience. Recruitment occurred between September 27, 2012 and October 2, 2012. Measures and procedures were identical to Study 3. As in Study 3, we contracted for 205 participants and SurveyMonkey collected 246 responses with loneliness and showering habit data (the company leaves the survey open to give their panelists time to respond to their requests). SurveyMonkey calculated the response rate at 29.6% and the average time to complete the survey was 4 minutes and 57 seconds. As before we only included the participants who answered "Yes" to the honesty question and correctly responded to both careless responding items (n = 228, 47.8% women; M age 48.23 years, SD = 15.31). This represents approximately 93% of the completed surveys. As with the other studies, participants largely self-identified as White (92.5%). Physical Warmth index was created by standardizing the three bathing/showering items and averaging them into a composite (alpha = .18).

Results and Discussion

There was no evidence for an association between Loneliness (M = 2.06; SD = .60, alpha = .95) and the Physical Warmth Index (r = -.10, p = .146, n = 228; 95% CI = -.23 to .03). The absolute size of the correlation in this study was almost as large as in Study 3, but it was in the opposite direction. The same null result was obtained when we used all participants with relevant data (r = -.07, p = .305, n = 246, 95% CI = -.19 to .06). Separate analyses with each shower/bath suggested that lonelier participants took fewer showers but the duration of their showers/baths

was longer (see Table 1). The hypothesis relevant correlation between the water temperature item and the loneliness scale was not statistically distinguishable from zero (r =.01, p = .883, n = 228, 95% CI = -.12 to .14). These results underscore the reality of sampling error when conducting research and should prove reassuring to readers concerned about the p = .063 result in Study 3.

Bayesian Analysis of the Combined Sample of 1,153 Participants

Kruschke (2011; 2013) argues that Bayesian estimation methods can provide a rigorous approach for determining whether researchers should accept the null hypothesis whereas such a determination is not strictly possible using the framework of traditional null hypothesis significance testing. Bayesian estimation procedures are used to produce a number of parameter estimates that could give rise to the observed data incorporating any pre-existing expectations about the parameter values (so-called priors). The relevant parameter for the current research is the regression coefficient linking loneliness and the physical warmth extraction variable or population correlation coefficient (also known as rho). Markov chain Monte Carlo (MCMC) methods were used to estimate a large number of possible parameter values that are consistent with the observed data (see Kruschke, 2013 for a clear description of the mechanics of Bayesian estimation, p. 576).

In terms of inferences about the plausibility of the null hypothesis, the strategy is to specify a region of practical equivalence or ROPE around the formal null hypothesis (i.e., H_o: rho = .00) to capture the idea that the association between two variables is practically indistinguishable from zero (i.e., the null hypothesis is re-expressed such that H_o represents the hypothesis that rho differs from zero in only a trivial way, if at all). In other words, the primary

task is to determine whether the vast majority of parameter estimates fall within the ROPE. In such instances, researchers have evidence in favor of the null hypothesis (see Kruschke, 2013). We selected a ROPE between -.09 to .09 for the standardized regression coefficient linking loneliness to the physical warmth index. This ROPE is based on the proposal by Cohen (1988) that correlations of around |.10| are small and thus our interval captures parameter values that reflect a smaller than a "small" association. Kruschke (2013) suggests a generic ROPE of -.10 to .10 (see p. 577).

We used the default vaguely informed prior to estimate the parameter of interest using a data file that contained responses from our combined sample of 1,153 participants from Studies 1 to 4. This file is uploaded as part of the supplemental materials. Using scripts for R (R Development Core Team, 2012) developed by Kruschke (2011), we implemented MCMC methods to generate a set of 100,000 credible parameter values for the regression equation. The R script outputs a histogram of credible parameter estimates and reports the interval that contains 95% of the credible estimates for the standardized regression parameter (this is called the highest density interval or HDI). In this case, the mean estimate was -.03 and the 95% HDI was -.09 to .03 (84.8% of the credible values were below 0). The HDI falls within our ROPE (-.09 to .09) suggesting that these data are consistent with the null hypothesis. As a point of comparison, the standardized regression coefficient using OLS regression on this combined dataset of 1,153 was -.03 (p = .309; r = -.03, 95% CI = -.09 to .03). In short, there is very little reason to believe that the association between loneliness and the physical warmth index is different from zero given our data (see also the last column of Table 1). This was also true for the hypothesis relevant correlation between the water temperature item and the loneliness scale using both Bayesian and

traditional estimation methods (mean estimate = -.01; 95% HDI = -.07 to .04 with 68.2% of the values below 0; r = -.01, p = .632, n = 1152, 95% CI = -.07 to .04).

In sum, we found no support for the Bargh and Shalev results in Studies 1 to 4 when considering the physical warmth composite or the hypothesis relevant association between temperature and loneliness. None of the individual p values for the composite variable or the warmth variable were statistically significant at $p < .05$ and the relevant effect sizes were smaller than the original report. One concern with Studies 1 to 4 was that we used slightly different measures than were used by Bargh and Shalev (2012). Thus, we conducted five additional studies to address the possibility that slight measurement differences might have impaired our ability to duplicate the original results.

Study 5: Participants from Amazon.com's Mechanical Turk

Sample and Procedure

As in Study 2, participants were "workers" recruited through Amazon.com's Mechanical Turk. Recruitment occurred between October 8, 2012 and October 9, 2012. We randomly assigned participants to either complete the bathing/showering items first or the loneliness questions first. Participants then completed demographic items. We included a directed response item to detect careless responding embedded in the demographics questionnaire that asked participants to select the "rarely" response option. (We purposely did not embed an item within the lifestyle questionnaire to avoid modifying the original Bargh and Shalev questionnaire). We also included an item at the end of the demographic questions: "I responded to this survey honestly." The 494 participants who answered "Yes" to the honesty question and

responded correctly the "rarely" item were included in the current analyses (31.6% women; M age = 27.56 years, SD = 8.86). This represents approximately 98% of the data collected from mTurk for this study. The physical warmth index was created by standardizing the three bathing/showering items (see appendix) and averaging them into a composite (alpha = .07) after the frequency and temperature items were reverse coded so that higher scores indicate more frequent baths/showers and warmer baths/showers.

Results and Discussion

There was a statistically significant association between Loneliness (M = 2.21; SD = .69, alpha = .91) and the Physical Warmth Index (r = .10, p = .029, n = 494; 95% CI = .01 to .18). The same pattern was obtained when we used all participants with relevant data (r = .10, p = .029, n = 495, 95% CI = .01 to .18). Separate analyses with each shower/bath item suggested that lonelier participants took longer showers/baths (see Table 2). On the other hand, the most hypothesis relevant correlation between the water temperature item and the loneliness scale was not statistically distinguishable from zero (r =.06, p = .159, n = 493, 95% CI = -.02 to .15). The ordering of survey questions did not moderate the reported associations (Lifestyle First: r = .11, p = .084, n = 249; Loneliness First: r = .09, p = .178, n = 245; z for difference = 0.27) or generate mean-level differences in the two variables (minimum p = .309).

In short, Study 5 provided some evidence for the substitutability hypothesis given that the correlation between the composite variable and loneliness was statistically significant at $p < .05$. However, the .10 observed correlation would not have been statistically distinguishable from zero using the Bargh and Shalev sample sizes (e.g., p = .49 if n = 51 and r = .098). The .10 correlation was not statistically different from the .37 correlation in Study 1b of Bargh and

Shalev (z for difference = 1.73; 95% CI for the .27 difference = -.55 to .04), although the .10 correlation was statistically different from the .57 correlation in Study 1a (z for difference = 3.63, 95% CI for .47 difference =.25 to .69). One possibility is that the loneliness measure used by Bargh and Shalev was more sensitive to the showering/bathing effects than the complete 20-item UCLA Loneliness scale. Thus, we conducted a sixth study to investigate this possibility by administering both loneliness questionnaires in a random order to participants. Study 6 also provided insight about the convergence between the two measures of loneliness.

Study 6: Participants from Amazon.com's Mechanical Turk

Sample and Procedure

Participants were "workers" recruited through Amazon.com's Mechanical Turk. Recruitment occurred between October 9, 2012 and October 16, 2012. We randomly assigned participants to one of four conditions: (1) Lifestyle questions, Bargh and Shalev Loneliness scale, UCLA Loneliness scale; (2) Bargh and Shalev Loneliness scale, Lifestyle questions, UCLA Loneliness scale; (3) UCLA Loneliness scale, Lifestyle questions, Bargh and Shalev Loneliness scale; (4) Lifestyle questions, UCLA Loneliness scale, Bargh and Shalev Loneliness scale. Participants then completed demographic items. As in Study 5, we included a directed response item to detect careless responding embedded in the demographics questionnaire that asked participants to select the "rarely" response option. We also included an item at the end of the demographic questions: "I responded to this survey honestly." Only the 553 participants who answered "Yes" to the honesty question and responded correctly the "rarely" item were included in the current analyses (43.3% women; M age = 30.65 years, SD = 11.68). This represents approximately 97% of the data collected from mTurk for this study. The physical warmth index

was created by standardizing the three bathing/showering items and averaging them into a composite (alpha = -.05) after the appropriate items were reverse scored.

Results and Discussion

There was no evidence for an association between the Bargh and Shalev Loneliness scale (M = 2.16; SD = .73, alpha = .92) and the Physical Warmth Index (r = .08, p = .058, n = 553; 95% CI = .00 to .16) using the conventional alpha level of .05. There was no evidence for an association between the UCLA Loneliness scale (M = 2.29; SD = .61, alpha = .95) and the Physical Warmth Index (r = .03, p = .488, n = 553; 95% CI = -.05 to .11). Similar results were obtained when we used all participants with relevant data (Bargh and Shalev Loneliness Scale: r = .08, p = .057, n = 569, 95% CI = .00 to .16; UCLA Loneliness scale: r = .03, p = .508, n = 566, 95% CI = -.06 to .11). Separate analyses with each shower/bath item suggested that lonelier participants took longer showers/baths regardless of the measure of loneliness whereas lonelier participants also took fewer showers regardless of the measure. The specific warmth item was statistically significant for the Bargh and Shalev loneliness scale but not when considering the UCLA loneliness scale (p = .047 and .252, respectively); however, the two correlations were not significantly different from each other (z = 1.4; 95% CI for the difference: -.01 to .09).

The experimental manipulations did not have detectable effects at $p < .05$ on mean levels of the focal variables (the physical warmth index and two loneliness scales). The minimum p value was .133 in a series of three one way ANOVAs. We then used moderated multiple regression with three dummy codes for the four conditions to test whether condition impacted the associations between the loneliness measures and physical warmth variable. Condition did not moderate the association between the physical warmth index and the loneliness measures (range

of correlations for Bargh and Shalev measure: r = -.03 to .20; range of correlations for ULCA measures: r = -.11 to .12). Condition did not moderate the association between the two loneliness measures (range of correlations: r = .76 to .86; overall r = .82). In short, Study 6 did not provide strong evidence for the focal association between loneliness and warmth extraction from water. Nonetheless, we conducted three additional studies with college students to obtain more data on this topic. These serve as three replications of Study 1a in Bargh and Shalev (2012) using their same measures.

Study 7: College Students I

Sample and Procedure

Participants were 311 college students (68.8% women; M age = 19.69 years, SD = 2.26) who received course credit as part of the Michigan State University Psychology Subject Pool during the Fall Semester of 2012. Participants completed measures as part of a larger study designed to investigate recollections of mood on a previous day. The majority of participants self-identified as white (81.0%). Participants came to the lab and were greeted by trained research assistants. Participants provided informed consent and all survey items were administered via computers using a web-based survey interface.

The measures analyzed here were a subset of the individual difference measures completed by participants including a measure of the Big Five domains. Embedded within the Big Five personality questionnaire were two "instructed response" items and participants were also directed to select "rarely" as one of the demographic items. The last question of the entire survey asked students if they responded honestly. Only participants who correctly answered the

three instructed response items and indicated that they answered questions honestly were included in the current analyses (approximately 78.7% of the participants). The physical warmth index was created by standardizing the three bathing/showering items and averaging them into a composite (alpha = -.06) after the appropriate items were reverse scored.

Results and Discussion

There was no evidence for an association between the Bargh and Shalev Loneliness scale (*M* = 2.06; *SD* = .65, alpha = .90) and the Physical Warmth Index (*r* = .02, *p* = .719, *n* = 311; 95% CI = -.09 to .13). These same results were obtained when we used all participants with relevant data (*r* = .03, *p* = .541, *n* = 395, 95% CI = -.07 to .13). None of the separate analyses with each shower/bath item produced statistically significant correlations (minimum *p* = .055 for frequency). The hypothesis relevant correlation between the water temperature item and the loneliness scale was not statistically distinguishable from zero (*r* =.03, *p* = .644, *n* = 311, 95% CI = -.09 to .14).

Study 8: College Students II

Sample and Procedure

Participants were 365 college students (71.0% women; M age = 19.69 years, SD = 1.46) who received course credit as part of the Michigan State University Psychology Subject Pool during the Spring Semester of 2013. Participants completed measures as part of a larger study designed to investigate the personality correlates of expressive writing (i.e., the Spring 2013 parallel to Study 1). The majority of participants self-identified as white (75.5%). Participants came to the lab and were greeted by trained research assistants. Participants provided informed

consent and all survey items were administered via computers using a web-based survey interface.

As in Study 7, the measures analyzed here were a subset of the individual difference measures completed by participants including a measure of the Big Five domains. Embedded within the Big Five personality questionnaire were two "instructed response" items and the last question of the entire survey asked students if they responded honestly. Only participants who answered correctly to the two instructed response items and indicated that they answered questions honestly were included in the current analyses (approximately 81.1% of the participants). The physical warmth index was created by standardizing the three bathing/showering items and averaging them into a composite (alpha = .01) after the appropriate items were reverse scored.

Results and Discussion

There was no evidence for an association between the Bargh and Shalev Loneliness scale (M = 2.03; SD = .64, alpha = .90) and the Physical Warmth Index (r = .02, p = .766, n = 365; 95% CI = -.09 to .12). These same results were obtained when we used all participants with relevant data (r = .01, p = .902, n = 450, 95% CI = -.09 to .10). None of the separate analyses with each shower/bath item produced statistically significant correlations (minimum p = .061 for duration) with the caveat that the correlation between loneliness and duration was statistically significant using all participants (r = .10, p = .028, n = 450; 95% CI = .01 to .19). The hypothesis relevant correlation between the water temperature item and the loneliness scale was not statistically distinguishable from zero (r =.005, p = .925, n = 365, 95% CI = -.10 to .11).

Study 9: College Students III

Sample and Procedure

Participants were 197 college students (80.2% women; M age = 19.89 years, SD = 1.99) who received course credit as part of the Michigan State University Psychology Subject Pool during the Spring Semester of 2013. The majority of participants self-identified as white (80.7%). Participants were a subset of students enrolled in an unrelated experiment designed to test the impact of weather on mood. Participants were initially recruited into an experimental "pool" when they completed baseline measures of mood and subjective well-being along with our directed response items. Participants were then randomly assigned to be contacted on either a clear and sunny day or a relatively cold and grey day. Participants completed the loneliness and bath habits measures during the follow-up assessment after they completed the primary dependent measures of subjective well-being and mood. Only participants who correctly completed the two instructed response items and responded that they answered the baseline items honestly were included in these analyses. There was no evidence of an impact of weather condition on any of the loneliness and showering measures. The physical warmth index was created by standardizing the three bathing/showering items and averaging them into a composite (alpha = -.18) after the appropriate items were reverse scored.

Results and Discussion

There was no evidence for an association between the Bargh and Shalev Loneliness scale (*M* = 1.99; *SD* = .66, alpha = .92) and the Physical Warmth Index (r = -.130, p = .069, n = 197; 95% CI = -.26 to .01). These same results were obtained when we used all participants with

relevant data who did not pass the initial screening questions (r = -.079, p = .212, n = 251, 95% CI = -.20 to .05). None of the separate analyses with each shower/bath item produced statistically significant correlations (minimum p = .095 for the warmth item). The hypothesis relevant correlation between the water temperature item and the loneliness scale was not statistically distinguishable from zero (r = -.120, p = .095, n = 196, 95% CI = -.26 to .02). There were no mean level differences between conditions for loneliness scale or the bathing/showering composite (minimum p = .240). In general, the results of Study 9 were consistent with our three other college student samples.

Bayesian Analysis of the Combined Sample of 1,920 Participants Completing the Original Bargh and Shalev (2012) measures

As with the measures used in Studies 1 to 4, we conducted a Bayesian analysis on the combined sample of 1,920 participants from Studies 5 to 9 who completed the exact same measures as used by Bargh and Shalev (2012). This file is included as part of the supplemental materials associated with this article. The mean estimate for the standardized regression coefficient was .01 and the 95% HDI was -.03 to .06 (28.2% of the credible values were below 0). As a point of comparison, the standardized regression coefficient from OLS regression was .013 (p = .561, n = 1,920; r = .013; 95% CI: -.03 to .06). There was no evidence for any association between the water temperature item and the loneliness scale (mean estimate = .02; 95% HDI = -.02 to .06 with 19.2% of the values below 0; r =.02, p = .385, n = 1,916, 95% CI = -.02 to .06).

Local Meta-Analysis and Power

To provide an overall perspective on the association between loneliness and personal bathing habits, we conducted a local meta-analysis of our nine studies combined with Studies 1a and 1b in Bargh and Shalev (2012) using the Comprehensive Meta-Analysis 2.0.064 software package (Borenstein, Hedges, Higgins, & Rothstein, 2005). We computed z-scores for the two loneliness measures in Study 6 and averaged them to form a loneliness composite variable to compute summary effect sizes from Study 6. The results across our nine studies were generally consistent so we believe it is reasonable to aggregate across the exact and near-exact replication studies. Nonetheless, we used a random-effects model rather than a common effect (or fixed-effect) model (see Borenstein, Hedges, Higgins, & Rothstein, 2009). The individual studies used different measures and involved different samples (college students, community members from New Haven, internet panelists, and mTurk workers) so we did not want to make a strong assumption concerning a single overarching population effect size. Cumming (in press) recommends that researchers use the random effects meta-analytic model when faced with a choice between the random-effects and fixed-effect models. Results are displayed in Table 3 along with the 95% CI for the meta-analytic correlations.

There was no evidence that the meta-analytic correlation was statistically different from zero for the composite index or the theoretically-relevant warmth variable. The meta-analytic correlation was different from zero for the frequency and duration variables but the signs were in opposite directions. Loneliness was positively related to duration but negatively related to frequency. Similar results were obtained when we used a common effect (or fixed-effect) meta-analysis model as reported in Table 3 for interested readers.

We also considered the power of each of our studies to detect the original effects. In terms of the index composite variable, each study had over .99 power to detect a correlation of .37 (the smallest of the Bargh and Shalev effect sizes for the composite index variable). Each of our 9 studies had .96 power or higher to detect the smallest effect size for temperature (i.e., r = .26) and duration (r = .29) from the original Bargh and Shalev studies (see e.g., Table 3). However, our studies did not have adequate power (i.e., .80 or higher; Cohen, 1988, p. 56) to detect a correlation of .03 for the frequency correlation from their Study 1b (Min Power for Study 9 = .05; Max Power for Study 6 = .11). On the other hand, this is arguably a trivial correlation in the first place and we had over .97 power to detect the meta-analytic average of the frequency effect from their Study 1a and 1b (r = .275). In broader terms, none of our individual studies had adequate power to detect a small effect size (i.e., $r \geq |.10|$; Min Power for Study 9 = .29; Max Power for Study 6 = .65) but each study had over .90 power to detect a medium effect size (i.e., $r \geq |.30|$). A sample size of around 780 is required to detect a small correlation with .80 power and alpha set to .05 with a two-tailed test. Only the two aggregated files from Studies 1 to 4 and Studies 5 to 9 exceeded this requirement.

Post-Hoc Analyses

A number of analyses were suggested through the review process. These provide additional perspective on the underlying data. One issue concerned the use of the z-score composite for the loneliness measure from Study 6. Thus, we repeated the meta-analysis using the coefficients from Table 2 for the Bargh and Shalev (2012) loneliness measure. Results were similar to those reported in Table 3 with the caveat that the fixed-effect estimate for the index variable, though smaller in size than the random-effect estimate, passed the .05 alpha threshold

for statistical significance (Frequency: Random = -.082, 95% CI = -.116 to -.045, $p < .001$; Fixed = -.065, 95% CI = -.127 to -.002, $p = .044$; Warmth: Random = .028, 95% CI = -.017 to .072, $p = .222$; Fixed = .029, 95% CI = -.006 to .064, $p = .099$; Duration: Random = .114, 95% CI = .071 to .156, $p < .001$; Fixed = .115, 95% CI = .081 to .150, $p < .001$; Index: Random = .053, 95% CI = -.022 to .127, $p = .165$; Fixed = .036, 95% CI = .001 to .071, $p = .044$).

A second issue was whether the gender composition of the samples moderated effect sizes using meta-regression (see Borenstein et al., 2005). There were statistical indications at $p < .05$ that sample composition impacted effect size estimates for the warmth, duration, and index variables. In all three cases, samples with proportionately more women yielded more negative effect sizes. Graphs of these effects are available upon request. Sample level moderator effects for gender were not statistically detectable when considering frequency.

Given the meta-regression results for gender, we computed effect sizes separately for women and men in each of the 11 samples (a complete table available upon request) and we then meta-analyzed the individual estimates. Data on gender were missing from 2 participants in three of our studies (Study 2, Study 4, and Study 6). The estimates from both random-effects and fixed-effect models are reported in Table 4. In general, the effect size estimates were similar for women and men and similar to the results in Table 3. For example, the meta-analytic correlation for the index variable was .06 for women and .07 for men from the random-effects model. Neither of these coefficients was statistically distinguishable from zero. The meta-analytic correlation for the temperature variable was .03 for both women and men (both statistically indistinguishable from zero). The discrepancies in results when comparing the gender composition as a sample-level characteristic versus computing separate estimates for women and

men suggest that additional sample-level variable or variables correlated with the percentage of women in the sample might be producing differences in effect size estimates. This is a potentially important direction for the future.

Indeed, it is instructive to consider whether effect size estimates fluctuated across different sample-level characteristics. These analyses are limited by the relatively small pool of studies and thus all results should be viewed with considerable caution. Nonetheless, we tested whether the aggregated effect sizes from our nine studies differed from the aggregated effect sizes from the two Bargh and Shalev studies. There was statistically significant evidence of a difference for the frequency, warmth, and index variables at $p < .05$ using both fixed-effect and mixed-effects methods (see Borenstein et al., 2005). The contrast for the duration variable was $p = .059$ for comparisons based on the fixed-effect analysis and $p = .056$ for the mixed-effects model. Separate meta-analytic estimates are reported in Table 5 for these comparisons. The Bargh and Shalev effect sizes tended to be larger in absolute value than the estimates from our Studies 1 to 9 with the caveat the signs were different for the frequency variable.

In light of the fact that there were apparent differences between our studies and the Bargh and Shalev studies, we excluded their two studies from further analyses and tried to explain any heterogeneity within our nine studies. The Q-statistic was not statistically significant for the duration (10.825, $df = 8$, $p = .212$), warmth (6.983, $df = 8$, $p = .538$), and frequency (10.950, $df = 8$, $p = .205$) associations but it was significant for the index composite association (15.680, $df = 8$, $p = .047$) across our 9 studies. This suggested to us that these additional comparisons would not yield much insight, a fact that is underscored by the general similarity in estimates from the fixed-effect and random-effect models for our studies in Table 5. We found

no indications that effect size estimates differed for online studies versus data collected in the lab or for the mTurk samples versus the non-mTurk samples for the four associations we investigated. Likewise, there was no indication of a difference in effect size estimates when comparing Studies 1 to 4 versus Studies 5 to 9 (i.e., the sets of studies using slightly different measures). In short, our effect size estimates seemed fairly uniform across our 9 studies whereas there were more pronounced differences between the estimates from our studies versus the two Bargh and Shalev studies.

General Discussion

This set of nine studies tested the prediction that trait loneliness would be "positively associated with the frequency, duration, and preferred water temperatures" of showers and baths (Bargh & Shalev, 2012, p. 156). We found fairly consistent support for this prediction with respect to duration but not with respect to frequency or water temperature. We obtained aggregated effect sizes that were indistinguishable from zero when considering the warmth variables directly (i.e., the most relevant correlation for the substitutability hypothesis) whereas the aggregated association for frequency variables were in the opposite direction of the original predication. Trait loneliness appears to be positively correlated with the duration of showers or baths but negatively correlated with the frequency of showering or bathing. The overall connection for the composite index variable and trait loneliness was close to zero.

We have no way of knowing the true population values for the associations we considered in this report and a number of unknown factors may explain discrepancies between the current results and the original Bargh and Shalev (2012) results. Additional research is needed to determine the precise magnitude of the effects in question and to identify variables that

might moderate associations between trait loneliness and showering and bathing habits. Indeed, we emphasize that future research is needed to explain fluctuations in effect size estimates across different samples with the caveat that the underlying effect sizes are likely to be small. We therefore suggest that new studies designed to test associations between trait loneliness and personal bathing habits plan to use extremely large sample sizes.

In sum, we doubt there is much practical or theoretical significance surrounding the empirical connection between trait loneliness and the tendency to take warm showers or baths. We hope it is uncontroversial to note there are substantial differences between correlations around .05 versus correlations around .50, both in terms of practical significance as well as theoretical meaningfulness. More broadly, we conclude that more research is needed before findings from Bargh and Shalev (2012) are used to inform treatment and intervention designed to ameliorate real-world loneliness.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denisseen, J. J. A., Fielder, K. et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108-119.

Bargh, J. A., & Shalev, I. (2012). The substitutability of physical and social warmth in daily life. *Emotion, 12,* 154-162.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43, 800-813.

Bonett, D. G., (2012). Replication-extension studies. *Current Directions in Psychological Science, 21,* 409-412.

Borenstein, M., Hedges, L., Higgins, J., Rothstein, H., (2005). Comprehensive Meta-Analysis (Version 2) [Computer Software]. Englewood, NJ: Biostat.

Borenstein, M., Hedges, l. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* London: John Wiley.

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives in Psychological Science, 6,* 3-5.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70,* 596-612.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Cumming, G. (in press). The new statistics: Why and how. *Psychological Science*.

Hemphill, J. F. (2003). Interpreting the magnitude of correlation coefficients. *American Psychologist, 58,* 78-80.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS.* New York: Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*, 573-603.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48,* 61-83.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44,* 1-23.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17,* 437-455.

R Development Core Team (2012). R: A language and environment for statistical computing. Retrieved from http://www.R-project.org.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331-363.

Russell, D. W. (1996). UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment, 66,* 20-40.

Russell, D. W., Peplau, L. A., & Ferguson, M. L. (1978). Developing a measure of loneliness. *Journal of Personality Assessment, 42,* 290-294.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17,* 551-566.

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

Simonsohn, Uri.(2013). Small telescopes: Detectability and the evaluation of replication results (December 10, 2013). Available at SSRN: http://ssrn.com/abstract=2259879 or http://dx.doi.org/10.2139/ssrn.2259879

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R. et al. (2011). Replication in prevention science. *Prevention Science, 12,* 103-117

Table 1: Correlations between Trait Loneliness and Bathing/Showering Items - Studies 1 to 4

	Frequency	**Temperature**	**Duration**	**Index**	**N**
Study 1	-.15*	-.03	.07	-.06	235
Study 2	-.11*	.02	.08	-.01	479-480
Study 3	.08	-.02	.17*	.13	209-210
Study 4	-.20*	.01	.02	-.10	227-228
Overall	-.12*	-.01	.08*	-.03	1,151 to 1,153
95% CI in Overall File	-.18 to -.06	-.08 to .05	.02 to .14	-.10 to .04	

Note: Frequency, Temperature, and Duration refer to the specific items included in the Physical Warmth Index. The exact item wording and corresponding frequency distributions for Studies 1 to 4 are reported in the appendix. The overall correlation was based on an aggregate data file that combined the data from Studies 1 to 4 into a single file. 95% Confidence Intervals constructed with Biased-Corrected Bootstrapping Procedures with 1,000 resamples using the aggregated data file.

* $p < .05$.

Table 2: Correlations between Trait Loneliness and Bathing/Showering Items - Studies 5 to 9

	Frequency	**Temperature**	**Duration**	**Index**	**N**
Study 5	-.07	.06	.18*	.10*	493-494
Study 6	-.10*	.09*	.16*	.08	551-553
Study 7	-.10	.03	.11	.02	311
Study 8	-.08	.01	.10	.02	365
Study 9	-.08	-.12	-.01	-.13	196-197
Overall	-.10*	.02	.10*	.01	1,916-1,920
95% CI	-.15 to -.05	-.03 to .07	.04 to .15	-.04 to .06	

Note: Frequency, Temperature, and Duration refer to the specific items included in the Physical Warmth Index. The exact item wording and corresponding frequency distributions for Studies 5 to 9 are reported in the appendix. The overall correlation was based on an aggregate data file that combined the data from Studies 5 to 9 into a single file. 95% Confidence Intervals constructed with Biased-Corrected Bootstrapping Procedures with 1,000 resamples using the aggregated data file. Correlations for the UCLA Loneliness measure in Study 6 were -.15, .05, .16, and .03, respectively.

* $p < .05$.

Table 3: Meta-Analytic Results

	Frequency	**Temperature**	**Duration**	**Index**
Study 1	-.15	-.03	.07	-.06
Study 2	-.11	.02	.08	-.01
Study 3	.08	-.02	.17	.13
Study 4	-.20	.01	.02	-.10
Study 5	-.07	.06	.18	.10
Study 6	-.13	.07	.16	.06
Study 7	-.10	.03	.11	.02
Study 8	-.08	.01	.10	.02
Study 9	-.08	-.12	-.01	-.13
Bargh and Shalev 1a	51	.26	.29	.57
Bargh and Shalev 1b	.03	.34	.33	.37
N	3,162	3,160	3,162	3,165
Random-Effects Model				
Point Estimate	-.068	.026	.115	.050
95% CI	-.132 to -.004	-.018 to .069	.071 to .158	-.024 to .123
p	.038	.245	.000	.186
Fixed-Effect Model				
Point Estimate	-.087	.027	.117	.032
95% CI	-.122 to -.053	-.008 to .062	.082 to .151	-.003 to .067
p	.000	.133	.000	.073

Note: Sample sizes of 51 and 41 were used for Bargh and Shalev Studies 1a and 1b, respectively. Effect sizes for Study 6 were calculated using the average of the z-scored UCLA Loneliness measure and the Bargh and Shalev Loneliness measure.

Table 4: Meta-Analytic Effect Size Estimates for Women and Men

	Frequency		**Temperature**		**Duration**		**Index**	
	Women	Men	Women	Men	Women	Men	Women	Men
Random-Effects Model								
Estimate	-.063	-.062	.044	.033	.114	.120	.056	.070
Lower Bound for 95% CI	-.128	-.159	-.036	-.020	.067	.041	-.031	-.041
Upper Bound for 95% CI	.001	.037	.123	.086	.161	.198	.142	.180
p	.055	.219	.284	.224	.000	.003	.210	.216
Fixed-Effect Model								
Estimate	-.069	-.109	.030	.033	.114	.121	.038	.027
Lower Bound for 95% CI	-.116	-.160	-.018	-.019	.067	.069	-.009	-.026
Upper Bound for 95% CI	-.021	-.056	.078	.085	.161	.172	.086	.079
p	.005	.000	.218	.216	.000	.000	.115	.315
Sample Size	1,563	1,357	1,560	1,359	1,561	1,361	1,562	1,360

Table 5: Meta-Analytic Comparisons Between Studies 1 to 9 versus Bargh and Shalev Studies 1a and 1b

	Frequency	**Temperature**	**Duration**	**Index**
Studies 1 to 9				
Fixed-Effect Model	-.098 (-.133 to -.063)*	.019 (-.017 to .054)	.111 (.076 to .146)*	.018 (-.018 to .053)
Random-Effects Model	-.097 (-.139 to -.055)*	.019 (-.017 to .054)	.107 (.066 to .149)*	.011 (-.040 to .061)
Bargh and Shalev				
Fixed-Effect Model	.296 (.094 to .475)*	.296 (.093 to .475)*	.308 (.106 to .485)*	.488 (.311 to .632)*
Random-Effects Model	.275 (-.198 to .644)	.296 (.093 to .475)*	.308 (.106 to .485)*	.484 (.269 to .654)*

Note: 95% CIs reported inside parentheses. * $p < .05$

Appendix: Responses to Showering/Bath Items

Studies 1 to 4 Frequency Question: ***How often do you usually take a bath/shower?***

Value	Response	Study 1	Study 2	Study 3	Study 4
1	Less than once a week	0.4%	0.6%	0.0%	0.9%
2	Once a week	0.4%	2.3%	2.9%	1.8%
3	Two-three times a week	3.8%	9.8%	7.2%	7.0%
4	Once every other day	17.4%	19.2%	14.4%	18.9%
5	Once a day	67.7%	57.5%	70.8%	63.2%
6	Two times a day	9.8%	10.0%	4.8%	8.3%
7	Three times a day	0.4%	0.2%	0.0%	0.0%
8	More than three times a day	0.0%	0.4%	0.0%	0.0%
	Sample Size	235	480	209	228

Studies 1 to 4 Temperature Question: ***How warm is the water you use when you take a bath/shower?***

Value	Response	Study 1	Study 2	Study 3	Study 4
1	Cold	0.0%	0.4%	0.0%	0.4%
2	Cool	0.4%	3.3%	1.0%	1.3%
3	Lukewarm	4.3%	6.9%	7.6%	4.4%
4	Warm	15.7%	24.2%	30.0%	24.1%
5	Very Warm	47.7%	35.3%	35.2%	38.6%
6	Hot	28.1%	27.3%	22.9%	30.3%
7	Very Hot	3.8%	2.5%	3.3%	0.9%
	Sample Size	235	480	210	228

Studies 1 to 4 Duration Question: ***When you do take a bath/shower, about how much time do you spend in the bath/shower?***

Value	Response	Study 1	Study 2	Study 3	Study 4
1	Less than 2 minutes	0.0%	0.6%	0.5%	0.0%
2	2-5 minutes	1.3%	6.5%	13.3%	10.6%
3	5-10 minutes	19.1%	26.5%	41.9%	40.5%
4	10-15 minutes	40.4%	37.4%	27.6%	30.0%
5	15-20 minutes	30.2%	20.3%	13.3%	15.4%
6	20-30 minutes	8.9%	8.8%	3.3%	3.5%
	Sample Size	235	479	210	227

Studies 5 to 9 Frequency Question: ***How often do you usually take a bath/shower?***

Value	Response	Study 5	Study 6	Study 7	Study 8	Study 9
1	More than 3 times a day	0.2%	0.0%	0.6%	0.8%	0.5%
2	3 times a day	0.6%	1.3%	0.3%	1.1%	0.0%
3	2 times a day	6.7%	9.6%	12.9%	11.5%	5.1%
4	Once a day	66.4%	63.9%	59.8%	66.0%	73.1%
5	Once every other day	19.0%	14.5%	22.8%	16.4%	17.3%
6	2-3 times a week	5.3%	7.8%	3.5%	3.6%	3.6%
7	Once a week	0.8%	2.2%	0.0%	0.3%	0.5%
8	Less than once a week	1.0%	0.7%	0.0%	0.3%	0.0%
	Sample Size	494	551	311	365	197

Studies 5 to 9 Temperature Question: ***What temperature do you use for the water when you take a bath/shower?***

Value	Response	Study 5	Study 6	Study 7	Study 8	Study 9
1	Very hot	6.1%	6.2%	8.7%	11.2%	13.8%
2	Hot	58.8%	57.5%	69.5%	62.2%	58.2%
3	Warm	28.4%	31.9%	20.9%	24.9%	24.5%
4	Lukewarm	5.3%	3.6%	1.0%	0.5%	2.6%
5	Cold	1.2%	0.7%	0.0%	1.1%	0.5%
6	Very Cold	0.2%	0.0%	0.0%	0.0%	0.5%
	Sample Size	493	551	311	365	196

Note: Response options in Bargh and Shalve (2012) are listed as ranging from cold to very hot (p. 156) but these options are consistent with the materials they provided to us.

Studies 5 to 9 Duration Question: ***About how much time do you spend in the bath/shower?***

Value	Response	Study 5	Study 6	Study 7	Study 8	Study 9
1	Less than 2 minutes	0.2%	0.2%	0.0%	0.0%	0.0%
2	2-5 minutes	8.1%	8.0%	2.3%	2.5%	1.5%
3	5-10 minutes	36.8%	36.3%	20.3%	21.1%	19.3%
4	10-15 minutes	34.8%	32.2%	36.7%	40.0%	35.0%
5	15-20 minutes	13.2%	13.7%	25.1%	23.6%	32.0%
6	20-30 minutes	5.1%	6.3%	12.2%	9.6%	9.1%
7	Over 30 minutes	1.8%	3.3%	3.5%	3.3%	3.0%
	Sample Size	494	553	311	365	197