Registered Replication Report: Srull & Wyer (1979)

Multilab direct replication of: Experiment 1 from Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology, 37,* 1660-1672.

Lead Authors:
McCarthy, Randy; Skowronski, John; Verschuere, Bruno; Meijer, Ewout; Ariane, Jim; Hoogesteyn, Katherine; Orthey, Robin

Contributing Authors:
Acar, Oguz A.; Aczel, Balazs; Bakos, Bence E.; Barbosa, Fernando; Baskin, Ernest; Bègue, Laurent; Ben-Shakhar, Gershon; Birt, Angie R.; Blatz, Lisa; Charman, Steve D.; Claesen, Aline; Clay, Samuel L.; Coary, Sean P.; Crusius, Jan; Evans, Jacqueline R.; Feldman, Noa; Ferreira-Santos, Fernando; Gamer, Matthias; Gerlsma, Coby; Gomes, Sara; González-Iraizoz, Marta; Holzmeister, Felix; Huber, Juergen; Huntjens, Rafaele J. C.; Isoni, Andrea; Jessup, Ryan K.; Kirchler, Michael; klein Selle, Nathalie; Koppel, Lina; Kovacs, Marton; Laine, Tei; Lentz, Frank; Loschelder, David D.; Ludvig, Elliot A.; Lynn, Monty L.; Martin, Scott D.; McLatchie, Neil M.; Mechtel, Mario; Nahari, Galit; Özdoğru, Asil A.; Pasion, Rita; Pennington, Charlotte R.; Roets, Arne; Rozmann, Nir; Scopelliti, Irene; Spiegelman, Eli; Suchotzki, Kristina; Sutan, Angela; Szecsi, Peter; Tinghög, Gustav; Tisserand, Jean-Christian; Tran, Ulrich S.; Van Hiel, Alain; Vanpaemel, Wolf; Västfjäll, Daniel; Verliefde, Thomas; Vezirian, Kévin; Voracek, Martin; Warmelink, Lara; Wick, Katherine; Wiggins, Bradford J.; Wylie, Keith; Yıldız, Ezgi

Proposing Researchers: Randy J. McCarthy & John J. Skowronski

Protocol vetted by: Robert Wyer

Protocol edited by: Daniel J. Simons

Citation: McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., …. Yildiz, E. (2017). Registered Replication Report: Srull & Wyer (1979). *Advances in Methods and Practices in Psychological Science*. Manuscript submitted for publication.

Address Correspondence to: Randy J. McCarthy, Center for the Study of Family Violence and Sexual Assault, Northern Illinois University; John J. Skowronski, Department of Psychology, Northern Illinois University. rmccarthy3@niu.edu

Acknowledgments: This project was partially supported by an NWO Replication Grant (401.16.001). Thanks to the Association for Psychological Science (APS) and the Arnold Foundation who provided funding to participating laboratories to defray the costs of running the study. Thanks to Robert Wyer for providing materials for the study and for providing guidance

about necessary changes to the protocol. Thanks to Katherine Wood for assistance with creating the forest plots.

Abstract

Srull and Wyer (1979) demonstrated that exposing participants to hostility-related stimuli caused them subsequently to interpret ambiguous behaviors as more hostile. In their Study 1, participants descrambled sets of words to form sentences. In one condition 80% of the descrambled sentences described hostile behaviors and in another condition 20% described hostile behaviors. All participants then read a vignette about a man named Donald who behaved in an ambiguously hostile manner and rated him on a set of personality traits. Next, participants rated the hostility of a list of ambiguously hostile behaviors (all on 0-10 scales). Participants who descrambled mostly hostile sentences rated Donald and the ambiguous behaviors as approximately three scale points more hostile than those who descrambled mostly neutral sentences. This Registered Replication Report describes the results of 26 independent replications (N = 7,373 in the total sample, k = 22 labs and N = 5,610 in the primary analyses) of Srull and Wyer (1979), each of which followed a pre-registered and vetted protocol. A random-effects meta-analysis showed the protagonist was seen as 0.08 scale points more hostile when primed with 80% hostile sentences than when primed with 20% hostile sentences (95% CI [0.004, 0.16]). Ratings of the ambiguously hostile behaviors were seen as 0.08 points less hostile when primed with 80% hostile sentences than when primed with 20% hostile sentences (95% CI [-0.18, 0.01]). Although the confidence interval for one outcome excluded zero and was in the predicted direction, these results suggest the currently-used methods do not produce an assimilative priming effect that is practically and routinely detectable.

Keywords: hostility, priming, impression formation, replication, ManyLabs, preregistration

Registered Replication Report: Srull & Wyer (1979)

In a now-classic study, Srull and Wyer (1979; SW hereafter) demonstrated that exposure to hostility-related stimuli affected how people subsequently interpreted the actions of a person described in a brief vignette and how they rated ambiguously hostile behaviors. SW has had considerable influence on the field of social cognition: SW is heavily cited, the "Donald" vignette has been used in several subsequent studies (e.g., Bartholow & Heinz, 2006; Devine, 1989; Philippot, Schwarz, Carrera, De Vries, & Van Yperen, 1991), the original findings have inspired many conceptual replications and extensions (e.g., Bargh & Pietromonaco, 1982; Herr, 1986; Mussweiler & Damisch, 2008), and SW is considered foundational both in the hostility priming literature and for studies that have extended priming effects beyond the domain of social judgments (e.g., Bargh, Chen, & Burrows, 1996; Dijksterhuis & van Knippenberg, 1998). A review and meta-analysis of this literature (DeCoster & Claypool, 2004) found a moderately sized effect reflecting the impact of priming on judgments about social targets (d = 0.35, 95% CI: [0.30, 0.41]).

However, in recent years, the robustness and replicability of some prominent social priming findings have been questioned (e.g., Cesario, 2014; Molden, 2014). Given its foundational role and continued citation as evidence of how priming can influence social judgments (e.g., Bargh, 2006, 2014; Higgins & Eitam, 2014; Strack & Schwarz, 2016), SW meets the Registered Replication Report (RRR) criterion of having high "replication value." The current RRR sought to estimate the magnitude and reliability of the hostile priming effects reported in SW through a series of independently conducted direct replications.

Original Hostility Priming Methods and Effects

The primary effect of interest in the current RRR is a phenomenon known as *assimilative*

priming: An effect in which exposure to prime-relevant stimuli causes subsequent judgments to incorporate *more* of the qualities of the primed construct.[1] In SW, exposure to more hostility-relevant stimuli caused participants to subsequently judge both a man named Donald and ambiguously hostile behaviors to be more hostile. SW tested two predictions for such assimilative priming effects. First, different amounts of "activation" of a primed mental representation (manipulated by exposing people to more or fewer of the priming stimuli) should be associated with the extent to which social judgments were affected. Second, the activation of primed mental representations should decay with the passage of time, thereby reducing the influence of the primes on subsequent social judgments.[2]

In Study 1 of SW (the focus of this RRR), participants first completed a sentence-descrambling task in which they underlined 3 of 4 words that could then be used to create a grammatically correct 3-word sentence (e.g., 'hand break his nose' which can form the sentence 'break his nose' or 'break his hand'). Different groups of participants completed sets of scrambled sentences that, when unscrambled, contained different proportions of hostile behaviors. After the sentence-descrambling task, participants were directed to a second researcher who was ostensibly conducting a different study. The "other study" consisted of three tasks. In the first task, participants read a vignette about a day in the life of a man named Donald who displayed a number of behaviors that were ambiguously hostile (e.g., 'Donald insisted that the waitress replace all the silverware because it was dirty'). They then rated Donald on twelve traits using a scale with anchors *0 = not at all* and *10 = extremely*. Six of these traits (i.e., hostile, unfriendly, dislikable, kind [reverse scored], considerate [reverse scored], and thoughtful [reverse scored]) were averaged to form an index of the extent to which Donald was perceived as "hostile." In the second task, participants rated the hostility of 15 individual behaviors (e.g.,

'Refusing to let a salesperson enter their house') using a scale with anchors *0 = not at all hostile* and *10 = extremely hostile*. Five behaviors were clearly hostile, five behaviors were clearly not hostile, and five behaviors were ambiguous with respect to hostility. Responses to the five ambiguously-hostile behaviors were averaged into an index of the extent to which the ambiguous behaviors were perceived to be hostile. Finally, participants estimated the co-occurrence of hostility with 11 other traits. However, the results from these co-occurrence ratings were not reported in SW and; thus, were not included in the current RRR.

The design of SW Study 1 included a number of between-subjects variables:

a) subjects de-scrambled a total of either 30 sentences or 60 sentences;

b) the scrambled sentence sets contained either 80% hostile sentences or 20% hostile sentences;

c) the "other study" tasks were completed either immediately after the descrambling task, after a 1-hour delay, or after a 24-hour delay; and

d) participants read one of two different versions of the Donald vignette.

A total of 96 participants completed SW Study 1, with 4 participants in each cell of the 2 × 2 × 3 × 2 between-participants factorial design. The relevant hypotheses tested in SW were that the more hostile sentences a participant descrambled, the more they would (a) view Donald as hostile and (b) view the ambiguously-hostile behaviors as hostile.[3]

The priming effect reported in SW was large. For the ratings of Donald, the mean difference between the two cells most comparable to the effect tested in this RRR (the 30 total trials-no delay conditions; see below for details) was approximately 3 scale points on an 11-point scale. For the ratings of the ambiguously-hostile behaviors, the mean difference of the two cells most comparable to the effect of this RRR also was approximately 3 scale points on an 11-point

scale. However, there may have been an error in the statistics reported in the original article (personal communication from Robert Wyer to Daniel Simons, August 22, 2016). The possibility of an erroneously-reported statistic is consistent with the fact that the standard deviations reported in a similar study (Srull & Wyer, 1980) were approximately 6 times as large and evinced a substantially smaller effect size than SW (see DeCoster & Claypool, 2004 for a detailed discussion). The uncertainty about the size and credibility of the original effect further motivates the need for a precise estimate of the size of these priming effects.[4]

Methods

Contributing Labs

The current RRR involved 26 total data collection sites. Data from these sites were collected between November 2016 and November 2017. The study materials, which were originally created in English, were translated into 8 different languages (13 labs used materials in English, 5 in German, 4 in Dutch, 1 in French, 1 in Hebrew, 1 in Hungarian, 1 in Portuguese, 1 in Swedish, and 1 in Turkish [*note*: 2 labs used 2 languages]).

Study Participants

Total sample sizes for each contributing lab ranged from 207 to 377 participants (total *n* before exclusions = 7,372; 2,147 men; 5,175 women, and 51 missing gender information; mean age = 20.77, *SD* = 2.90). Table 1 describes the demographics of each individual sample. Each contributing lab pre-registered their data collection stopping rules prior to beginning data collection.

Procedure

Participants completed the SW study as part of a packet that included other tasks (see Table 2). After providing consent, participants provided demographic information and then

completed the tasks for this study. The materials for SW always came immediately after the first demographic information and always came before any tasks for the other RRR (see below).

Participants first completed the sentence-descrambling task. In this task, participants viewed 30 groups of 4 words (e.g., *him yell swear at*) and were instructed to underline 3 words that created a grammatically-correct sentence (e.g., *yell at him* or *swear at him*).[5] Some of these 30 items could only be completed as sentences describing hostile behaviors and others could only be completed as non-hostile or neutral sentences. Participants were randomly assigned to one of two conditions: Mostly hostile sentences (24/30 or 80% yield hostile sentences) or mostly neutral sentences (6/30 or 20% yield hostile sentences). Participants then read the vignette and rated the protagonist of the vignette on the same traits using the same response scale (*0 = not at all* to *10 = extremely*) as in SW. Participants then viewed and rated the hostility of the same set of behaviors (with minor modifications described below) using the same response scale (0 = *not at all hostile* to 10 = *extremely hostile)* as in SW.

Thus, the design of the current RRR had one between-participants variable (i.e., 80% Hostile primes vs. 20% Hostile primes) and two separate dependent variables (hostility ratings of the described individual and average hostility ratings of the ambiguously hostile behaviors).

Known Differences Between the RRR Study and SW

The SW RRR was developed in parallel with the Mazar, Amir, and Ariely RRR (Verschuere et al., 2018). Both RRRs were developed to be combined into one data collection effort, which allowed both RRRs to be framed as a series of unrelated tasks. The SW RRR used the original materials whenever possible, including the Donald vignette, the ratings of Donald, and the ratings of the ambiguously hostile behaviors. However, we had to either re-create or modify some of the study materials and we had to modify some aspects of the procedure to

accommodate the constraints of the RRR. Our decisions around these modifications were driven by a goal to minimize the differences between the current RRR and SW and to maintain the theoretically necessary conditions for an assimilative priming effect to emerge. These modifications also were made in consultation with Dr. Wyer.

The original sentence-descrambling stimuli were unavailable, so the lead author generated and pretested new stimuli (https://osf.io/32pkz/) that were consistent with the description of the original stimuli. Further, in consultation with Dr. Wyer, we modified the pronouns in the original list of behaviors to make them gender neutral and to fix minor wording errors. Given that younger participants may be unfamiliar with the action of slamming a handset onto a receiver to hang up a phone, we also changed one of the listed behaviors that described "slamming down a phone" to "abruptly hanging up a phone." Finally, because the name "Donald" might activate unwanted associations with Donald Trump after the 2016 election in the United States, we changed the name of the protagonist of the vignette from Donald to Ronald.

The purpose of the current RRR is to replicate the assimilative priming effect from SW. To do so, rather than including all of the factors in the original 2 × 2 × 3 × 2 design, we focused on a comparison of two conditions from SW that showed a clear effect. Given that all variables in the original study were manipulated between groups, excluding those variables should not affect the primary outcome measure. Thus, for both practical reasons (no need to have participants return later) and because it showed strong priming effects in the original study, we chose to focus on the immediate testing condition. Specifically, all participants in the current RRR completed 30 priming task trials wherein half of the participants descrambled sentences of which 80% (i.e., 24/30) were hostile and the other half descrambled sentences of which 20% (i.e., 6/30) were hostile. All participants completed the ratings of Ronald and of the behaviors

immediately after the priming tasks. Though this design does not permit a full assessment of all variables (i.e., delay, number of priming sentences) manipulated by SW, the pair of conditions that we chose to include provides a replication of the assimilative hostile priming effect reported in SW.

To simplify the counterbalancing scheme for the combined RRR, we also used only one of the two vignettes from the original SW study. This required us to select which of the two vignettes to use in the RRR. One vignette was reported in the text of SW and the other vignette, which was used in SW but not reported in the text of SW, was provided by Dr. Wyer in preparation for the RRR. Given the possibility that cultural norms for hostility have changed since 1979, the lead author conducted a norming study (https://osf.io/32pkz/) to assess how hostile the two Donald vignettes were viewed in the absence of priming. The vignette ultimately used in the RRR was judged to be somewhat less hostile and evinced slightly more variable ratings than the one reported in SW. Given the results of this norming study, and in consultation with Dr. Wyer, we elected to use the vignette that was not included in the text of SW.

Finally, one consequence of the need to include this RRR project as part of a larger packet of tasks is a modification to the cover story. In SW, participants were asked to complete the sentence-descrambling task ahead of another unrelated study. In the current RRR, the sentence-descrambling task and ratings tasks were completed as part of a single administration in a large classroom setting. Further, although the tasks for the SW RRR always appeared first, the anticipation of additional and presumably unrelated tasks could have induced a different task-completion mindset (e.g. “I gotta move along fast to get this done”) than might have been present in SW. As the RRR was being developed, Dr. Wyer noted that these features were potentially meaningful departures from the conditions of the original study. However, we believe the spirit

of the original cover story is maintained: The packet was described as a collection of separate tasks on writing, memory, imagination, judgment, and problem solving, and the priming and outcome tasks are distinct enough that participants likely viewed them as unrelated. Finally, other studies have successfully used sentence-descrambling tasks to examine hostile attributions without using the procedures described in SW (e.g., Bargh, Chen, & Burrows, 1996; Crouch, Skowronski, Milner, & Harris, 2008; DeWall & Bushman, 2009; Srull & Wyer, 1980; Wann & Branscombe, 1990).

Pre-Specified Exclusions

Given that this study was conducted in conjunction with another RRR, additional inclusion criteria that were specific to that RRR applied to the current study as well. Individual participants were not included if they did not complete the critical items for the RRR or if they did not follow the study instructions. Finally, participants less than 18 years old or older than 25 years old (which was an exclusion criterion for the other RRR) or did not provide gender information were not included. Labs were not included if they did not collect a minimum of 100 included participants in each condition (see https://osf.io/9afwn/ for details of the exclusion criteria).

In total, four labs did not collect the minimum of 100 included participants in each condition. These labs were omitted from the primary analyses, but included in the ancillary analyses. Among the 22 labs that were included in the primary analyses, sample sizes ranged from 204 to 348 participants (1,626 men; 3,984 women; mean age = 20.30, *SD* = 1.82; see Table 1 for information about each individual lab). Disclosures about data collection, participant recruitment and compensation, and any departures from the overall protocol can be found at https://osf.io/hrju6/.

Results

All study materials can be found on the project's Open Science Framework page (https://osf.io/vxz7q/). All analyses were pre-registered and all analysis scripts were written before viewing any data from the RRR studies. Any deviations from the pre-registered analyses scripts are commented clearly in the post-data analysis scripts (the pre-data and post-data R scripts are available at https://osf.io/jp45u/). The following meta-analyses used a random-effects model and the restricted maximum-likelihood (REML) estimator for estimating the amount of heterogeneity, and were conducted using the metafor package in R (e.g., Viechtbauer, 2010).

Analyses of Primary Hypotheses

Judgments of Ronald's Hostility. As in SW, ratings of Ronald on the six traits—hostile, unfriendly, dislikable, kind, considerate, and thoughtful—were averaged (after reverse coding "kind," "considerate," and "thoughtful") to yield a hostility index score for each subject. We then obtained an average hostility rating for the 80% Hostile and 20% Hostile priming conditions for each lab. From these, we conducted a random-effects meta-analysis on the difference between conditions in the hostility index to obtain an overall estimate of the size of the hostility priming effect.

Based on the Figure 1 in SW, participants in the 80% Hostile priming condition rated Donald as approximately 3 scale units more hostile (on a 0-10 scale) than did those in the 20% Hostile priming condition. The meta-analysis of the 22 RRR studies that met our inclusion criteria of having at least 100 participants in each condition observed an overall difference of 0.08 points (95% CI [0.004, 0.16]). The heterogeneity of this effect across labs was no bigger than what would be expected due to sampling error alone ($\tau = 0.08$; $Q(21) = 25.31$, $p = .23$), and

I^2 indicated that about 17.73% of the observed variance between effect sizes was caused by systematic differences between studies.

Judgments of Ambiguously Hostile Behaviors. As in SW, we averaged each participant's hostility ratings for the five ambiguously hostile behaviors for the 80% Hostile and 20% Hostile priming conditions for each lab. These 5 ambiguously hostile behaviors were: *Telling a garage mechanic that they will have to go somewhere else if the mechanic cannot fix their car that same day; Refusing to let a salesperson enter their house; When asked to donate blood to the Red Cross, lying by saying they had diabetes and therefore could not do so; Demanding their money back from a sales clerk;* and *Refusing to pay their rent until the landlord paints their apartment*. From these, we conducted a random-effects meta-analysis on the difference between conditions in the hostility ratings to obtain an overall estimate of the size of the hostility priming effect.

Based on the Figure 2 in SW, participants in the 80% Hostile priming condition of SW rated the ambiguous behaviors as approximately 3 scale units more hostile (on a 0-10 scale) than did those in the 20% Hostile priming condition. The meta-analysis of the 22 RRR studies that met our inclusion criteria of having at least 100 participants in each condition observed a difference of -0.08 points (95% CI:[-0.18, 0.01]). The heterogeneity of this effect across labs was no bigger than what would be expected due to sampling error alone ($\tau = 0.10$, $Q(21) = 24.39$, $p = .27$), and I^2 indicated that about 18.03 % of the observed variance between effect sizes was caused by systematic differences between studies.

Ancillary Analyses

We conducted two sets of ancillary analyses. The first examined the pattern of results when including all laboratories and participants regardless of the size of the final sample. A second set examined whether the language of the stimuli moderated these effects.

The impact of exclusion criteria. The primary analyses excluded data from laboratories that contributed fewer than 100 participants in each priming condition. The first ancillary analysis included data from all laboratories and participants even if they did not meet that criterion. Note that the exclusion criteria for individual participants (e.g., completing all priming trials, reporting demographic information, etc.) were still applied in this analysis.

In this full sample, which included 26 labs with 6,404 total participants, we observed a difference of 0.07 for the trait ratings of Ronald (95% CI [0.003, 0.14]; see Figure 3) and a difference of -0.10 for the behavior ratings (95% CI [-0.19, -0.001]; see Figure 4). For the trait ratings of Ronald, the heterogeneity of this effect across labs was no bigger than what would be expected due to sampling error alone, $\tau = 0.05$, $Q(25) = 25.89$, $p = .41$, $I^2 = 7.10$. For the behavior ratings, the heterogeneity of this effect across labs was no bigger than what would be expected due to sampling error alone, $\tau = 0.13$, $Q(25) = 35.03$, $p = .09$, $I^2 = 28.86$.

Overall, the results with the full sample were nearly identical to the results based on labs with more than 100 participants per condition.

Moderation by language. The original stimuli were created in English. We examined whether the language of the lab moderated the effect. Two labs administered the study using both a non-translated version and a translated version, which effectively allowed us to compute an effect for each version from these labs. Thus, the following analyses included 28 effects (i.e., 26 labs of which 2 provided two effects). The original English version of the study was used in 13

samples and these stimuli were translated into eight separate languages (German, $k = 5$; Dutch, $k = 4$; French, $k = 1$; Hebrew, $k = 1$; Hungarian, $k = 1$; Portuguese, $k = 1$; Swedish, $k = 1$; and Turkish, $k = 1$). For purposes of the moderation analysis, we tested whether effects from the translated versions (regardless of the translation) differed from the effects from the non-translated (i.e., English) versions. Thus, the non-translated vs. translated comparison had 1 degree of freedom.

For the trait ratings of Ronald, effects from the translated versions of the stimuli were not significantly different than the non-translated, English version, $QM(1) = 0.12$, $p = .73$. For the ratings of the ambiguous behaviors, effects from the translated versions of the stimuli were not significantly different than the non-translated, English version, $QM(1) = 1.36$, $p = .24$.

Discussion

In recent years, the replicability of assimilative priming effects has come into question. Results reported by other RRRs (e.g., Cheung et al., 2016; O'Donnell et al., 2017), ManyLabs studies (e.g., Klein et al., 2014), and individual studies (e.g., Doyen, Klein, Pichon, & Cleeremans. 2012; McCarthy, 2014; Pashler, Coburn, & Harris, 2012) have not found evidence of such priming effects. This context of doubt provided a reason to explore the replicability of one of the most influential assimilative priming effects in the field of social cognition: The hostile priming effect reported in SW.

The current RRR had two outcome variables. For the first outcome, participants who completed 80% hostile primes—the group theorized to be *more* primed by hostility—rated the protagonist in an ambiguously hostile vignette to be 0.08 points more hostile on a 0-10 scale than did participants who completed 20% hostile primes. The 95% confidence interval around this estimate excluded zero (i.e., the meta-analytic assimilative priming effect was significantly

different from zero), and 18 of the 26 labs produced an effect that was numerically in the predicted direction. However, the overall effect was much smaller than both the original effect reported in SW and the expected effect size derived from reviews of the published literature (e.g., the DeCoster & Claypool [2004] meta-analysis).

For the second outcome, participants who completed 80% hostile primes rated ambiguously hostile behaviors as 0.08 points on a 0-10 scale *less* hostile than did participants who completed 20% hostile primes. Not only is this effect smaller than the original effect reported in SW, it is numerically in the opposite direction. Only 9 of the 26 labs produced an effect in the predicted direction. In short, the meta-analytic effects of assimilative priming for both outcome measures were close to 0 scale units, a much smaller differences than the approximately 3-scale-unit differences reported by SW.

One possible explanation for the discrepancies between the RRR results and previously reported effects is that the published literature exhibits publication bias that leads to an inflated view of the magnitude and replicability of the SW hostility priming effect. Indeed, in the DeCoster and Claypool (2004) meta-analysis, there is a negative relationship between the magnitude of the published effects and the precision of those effects, a pattern that is consistent with (but not definitive proof of) the presence of publication bias. In the presence of publication bias, the literature might paint a misleading picture of the replicability and magnitude of assimilative priming effects. Unsurprisingly then, when publication bias is eliminated from the data, as in the current RRR, the obtained effect size is much smaller than a simple synthesis of the published literature would suggest.

Method differences between SW and the RRR also might contribute to their discrepant results. In comparison to the SW study, the RRR used different sentence-descrambling primes,

only one of the two original SW vignettes, and a different name for the protagonist (Ronald rather than Donald). Although such procedural details, either individually or in combination, could change the outcome of a study, it is hard to construct a cogent explanation for how they could do so. Moreover, we pretested the priming stimuli and the vignette to ensure that they activated the relevant constructs, and there is no obvious reason to believe the protagonist's name or other procedural differences should matter for obtaining an assimilative priming effect.

However, other differences in the SW and RRR methods might more plausibly contribute to between-study differences in outcomes. In SW, participants were exposed to an unexpected task (the sentence-descrambling task) before completing the task for which they had signed up (supposedly unrelated to the sentence-descrambling task). In the RRR, both the priming task and the person judgment task were framed as unrelated, but both appeared as a part of a lengthy booklet. This difference in cover story could have led to different results. For example, the booklet length could have induced a task-completion mindset (e.g. "I gotta move along fast to get this done") that might not have been present in SW, leading to shallower stimulus processing than in the original. The group context also might have led RRR participants to be less attentive to the study materials, and assimilative priming effects might be weakened as a result. During the planning phase of the project, Dr. Wyer noted this change in the cover story as a possible reason to expect a different outcome. However, Srull and Wyer (1980), which included conditions that replicated the assimilative priming effects in SW, used a procedure that involved only one researcher who gave participants a study packet containing "a wide array of experiments, contributed by various members of the psychology faculty, over the course of 2 hours" (p. 845). In the 1980 publication, Srull and Wyer justify this procedural choice by stating that "these instructions, along with the fact that the tasks were highly dissimilar, were intended to make

subjects think there was no relationship between any two tasks in the sequence" (p. 845). Given this precedent, it seems that neither the single experimenter nor the lengthy packet of "unrelated" tasks has historically been considered a barrier to creating the conditions necessary to produce an assimilative priming effect.

We also can exclude one procedural difference as a plausible explanation for the different outcomes. Several labs translated their priming task materials into non-English languages, and priming effects might hinge on subtle differences in meaning despite quality controls for these translations. However, the effects were generally homogenous across labs, so the language difference does not appear to explain the effect size difference.

In sum, we observed a small assimilative priming effect in the predicted direction for ratings of Ronald (i.e., the confidence interval for ratings of Ronald excluded zero) and a similarly small effect in the opposite direction for judgments about behaviors. Both effect size estimates were close to zero and were substantially smaller than those previously reported in published research. Our results suggest the procedures used in the RRR are unlikely to produce an assimilative priming effect that researchers could practically and routinely detect. Indeed, to study priming effects as small as the 0.08 scale unit difference we observed (which works out to approximately $d = 0.06$, 95% CI [0.01, 0.12]), a study would need 4,362 participants in each priming condition to have 80% power with an alpha set to .05. Although the current procedures were unfavorable for producing assimilative priming effects, other procedures, such as within-participants repeated-measures designs with a brief delay between the priming stimuli and the outcome measure, might provide a more promising approach for future assimilative priming research (e.g., Fazio, Jackson, Dunton, & Williams, 1995; Payne, Brown-Iannuzzi, & Loersch, 2016; Payne, Cheng, Govorun, & Stewart, 2005).

References

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91,* 340-345. doi:10.1080/00223890902935878

Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology, 36,* 147-168. doi:10.1002/ejsp.336

Bargh, J. A. (2014). The historical origins of priming as the preparation of behavioral responses: Unconscious carryover and contextual influences of real-world importance. *Social Cognition, 32,* 209-224. doi:10.1521/soco.2014.32.supp.209

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71,* 230-244. doi:10.1037/0022-3514.71.2.230

Bargh, J. A., Lee-Chai, A., Barndollar, K., Gollwitzer, P. M., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology, 81,* 1014-1027. doi:10.1037/0022-3514.81.6.1014

Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology, 43,* 437-449. doi:10.1037/0022-3514.43.3.437

Bartholow, B. D., & Heinz, A. (2006). Alcohol and aggression without consumption alcohol cues, aggressive thoughts, and hostile perception bias. *Psychological Science, 17,* 30-37. doi:10.1111/j.1467-9280.2005.01661x

Bless, H., & Schwarz, N. (2010). Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. In M.P. Zanna (Ed.), *Advances in experimental social psychology* (Vol 42, pp. 319-373). San Diego, CA, US: Academic Press.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9,* 40-48. doi:10.1177/1745691613513471

Chabris, C.F., Engel, D., Kim, Y.J., Loken, E., Woolley, A.W., Malone, T.W., et al. (2018). *Using collective intelligence to develop a new test of individual intelligence.* Manuscript in preparation.

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., ... & Carcedo, R. J. (2016). Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*, 750-764. doi:10.1177/1745691616664694

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83,* 1314–1329. doi:10.1037/0022-3514.83.6.1314

Crouch, J. L., Skowronski, J. J., Milner, J. S., & Harris, B. (2008). Parental responses to infant crying: The influence of child physical abuse risk and hostile priming. *Child Abuse & Neglect, 32,* 702-710. doi:10.1016/j.chiabu.2007.11.002

DeCoster, J., & Claypool, H. M. (2004). A meta-analysis of priming effects on impression formation supporting a general model of informational biases. *Personality and Social Psychology Review, 8,* 2-27. doi:10.1207/S15327957PSPR0801_1

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5-18. doi:10.1037/0022-3514.56.1.5

DeWall, C. N., & Bushman, B. J. (2009). Hot under the collar in a lukewarm environment: Words associated with hot temperature increase aggressive thoughts and hostile perceptions. *Journal of Experimental Social Psychology, 45,* 1045-1047. doi:10.1016/j.jesp.2009.05.003

Dijksterhuis, A., & van Kinppenberg, A. (1998). The relation between perception and behavior, or how to win a game of Trivial Pursuit. *Journal of Personality and Social Psychology, 74,* 865-877. doi:10.1037/0022-3514.74.4.865

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLOS ONE, 7,* e29081.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013–1027. doi:10.1037/0022-3514.69.6 .1013

Guilford, J. P. (1967). The nature of human intelligence. New York: McGraw-Hill

Herr, P. M. (1986). Consequences of priming: Judgment and behavior. *Journal of Personality and Social Psychology, 51,* 1106-1115. doi:10.1037/0022-3514.51.6.1106

Herr, P. M., Sherman, S. J., & Fazio, R. H. (1983). On the consequences of priming: Assimilation and contrast effects. *Journal of Experimental Social Psychology, 19,* 323-340. doi: 10.1016/0022-1031(83)90026-4

Higgins, E. T., Bargh, J. A., & Lombardi, W. (1985). Nature of priming effects on categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11,* 59-69. doi/10.1037/0278-7393.11.1.59

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology, 45,* 132-142. doi:10.1027/1864-9335/a000178

Martin, L. L. (1986). Set/reset: Use and disuse of concepts in impression formation. *Journal of Personality and Social Psychology, 51,* 493-504. doi:10.1037/0022-3514.51.3.493

McCarthy, R. J. (2014). Close replication attempts of the heat priming-hostile perception effect. *Journal of Experimental Social Psychology, 54,* 165-169. doi:10.1016/j.jesp.2014.04.014

McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Profile of Mood States Manual.* San Diego, CA: Educational and Industrial Testing Service.

Molden, D. C. (2014). Understanding priming effects in social psychology: What is “social priming” and how does it occur. *Social Cognition, 32,* 1-11. doi:10.1521/soco.2014.32.supp.1

Mussweiler, T., & Damisch, L. (2008). Going back to Donald: How comparisons shape judgmental priming effects. *Journal of Personality and Social Psychology, 95,* 1295-1315. doi:10.1037/a0013261

O’Donnell, M., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S. … Zrubka, M. (2018). Registered Replication Report: Dijksterhuis & van Knippenberg (1998). *Perspectives on Psychological Science*. doi:10.1177/1745691618755704

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLOS ONE*, *7*, e42510.

Payne, B. K., Brown-Iannuzzi, J. L., & Loersch, C. (2016). Replicable effects of primes on human behavior. *Journal of Experimental Psychology: General*, *145*, 1269-1279. doi:10.1037/xge0000201

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89,* 277–293. Doi:10.1037/0022-3514.89.3.277

Philippot, P., Schwarz, N., Carrera, P., De Vries, N., & Van Yperen, N. W. (1991). Differential effects of priming at the encoding and judgment stage. *European Journal of Social Psychology, 21,* 293-302. doi:10.1002/ejsp.2420210403

Rivers, A. M., & Sherman, J. (2018, January 19). *Experimental design and the reliability of priming effects: Reconsidering the "train wreck"*. Retrieved from psyarxiv.com/r7pd3

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology, 37,* 1660-1672. doi:10.1037/0022-3514.37.10.1660

Srull, T. K., & Wyer, R. S. (1980). Category accessibility and social perception: Some implications for the study of person memory and interpersonal judgments. *Journal of Personality and Social Psychology, 38,* 841-856. doi:10.1037/0022-3514.38.6.841

Strack, F. & Schwarz, N. (2016). Social priming-information accessibility and its consequences. *Current Opinion in Psychology, 12,* iv-vii. doi:10.1016/j.copsyc.2016.11.001

Verschuere, B., Meijer, E. H., Hoogesteyn, K..,McCarthy, R., Skowronski, J., …. (2018). Registered Replication Report: Mazar, Amir, & Ariely (2008). *Advances in Methods and Practices in Psychological Science*. Manuscript submitted for publication.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1-48.

Wann, D. L., & Branscombe, N. R. (1990). Person perception when aggressive or nonaggressive sports are primed. *Aggressive Behavior, 16,* 27-32. doi:10.1002/1098-2337(1990)16

Footnotes

[1] There also are *contrastive priming effects* wherein increasing exposure to priming stimuli causes judgments that social targets have *less* of the quality of the primed construct (e.g., Bless & Schwarz, 2010; Martin, 1986). For example, a contrastive hostile priming effect would be when exposure to hostile primes causes subsequent judgments that a social target is *less* hostile (e.g., Herr, 1986).

[2] However, the prediction that the influence of the prime will weaken over time is not a given. For example, some researchers have supposedly primed goals, which theoretically involve auxiliary cognitive processes that can maintain or even increase the effect of the priming stimuli on outcome variables with the passage of time (e.g., Bargh, Lee-Chai, Barndollar, Gollwitzer, & Trötschel, 2001).

[3] The logistics of the current RRR precluded us from manipulating the delay between the priming task and the social judgment tasks. Thus, the current RRR did not include any of the delay conditions that were included in SW.

[4] Notably, Study 2 of SW conceptually replicated the hostility priming findings (with somewhat weaker effects) by assessing the impact of "kindness" priming on social judgments of kindness. However, the RRR focuses only on the hostility priming result.

[5] Some labs reported difficulty when literally translating each word of the sentence-descrambling task from English into other languages (e.g., issues with gendered words or the way articles are used). In some cases, to allow for successful translations, the option words were changed slightly or the instructions were changed so that participants unscrambled "4 words or phrases." See individual labs' translations for details, (https://osf.io/hrju6/wiki/home/).

Table 1.
Demographics for individual labs

	Full Sample				Sample After Exclusions[b]		
Lab	*N*	Male/Female (missing)	Age(*SD*)	Included in Primary Analyses[a]	*N*	Male/Female	Age(*SD)*
Acar	237	82/153/2	21.15(2.03)	Yes	214	76/138	20.96(1.58)
Aczel	245	53/191/1	20.82(1.73)	Yes	225	47/178	20.76(1.63)
Baskin	207	105/102/0	19.63(0.90)	No	198	99/99	19.60(0.79)
Birt	234	46/188/0	21.50(4.52)	Yes	205	37/168	20.37(2.09)
Blatz	320	48/264/8	22.05(3.58)	No	212	24/188	20.66(2.19)
Evans	332	97/234/1	21.68(3.20)	Yes	243	69/174	20.94(1.68)
Ferreira-Santos	291	76/214/1	19.99(4.34)	Yes	234	59/175	19.35(1.60)
González-Iraizoz	235	39/196/0	18.65(0.88)	Yes	229	38/191	18.64(0.87)
Holzmeister	274	130/143/1	21.89(2.13)	Yes	253	118/135	21.62(1.61)

Huntjens	216	62/152/2	20.85(2.06)	No	190	54/136	20.64(1.77)
klein Selle & Rozmann	337	76/258/3	22.29(1.72)	Yes	299	65/234	22.21(1.52)
Koppel	263	119/143/1	22.03(2.20)	Yes	242	108/134	21.76(1.73)
Laine	313	41/269/3	19.39(2.14)	Yes	253	32/221	19.24(1.31)
Loschelder	248	83/156/9	21.30(2.00)	Yes	226	79/147	21.13(1.63)
McCarthy	318	123/193/2	21.41(2.95)	Yes	279	106/173	20.88(1.66)
Meijer	377	97/279/1	20.31(1.90)	Yes	348	86/262	20.20(1.59)
Özdoğru	365	42/323/0	20.27(2.63)	Yes	332	36/296	19.96(1.32)
Pennington	255	51/196/8	20.29(4.44)	Yes	217	45/172	19.31(1.40)
Roets	253	28/224/1	18.44(2.02)	Yes	204	23/181	18.47(0.96)
Suchotzki	256	46/207/3	20.35(1.68)	Yes	246	44/202	20.30(1.65)
Sutan	304	154/148/2	20.64(0.91)	Yes	252	129/123	20.62(0.93)

Tran	277	77/200/0	24.59(3.55)	No	194	38/156	22.95(1.36)
Vanpaemel	288	64/224/0	20.27(3.16)	Yes	237	48/189	20.25(1.76)
Verschuere	302	88/213/1	19.76(2.20)	Yes	285	83/202	19.60(1.62)
Wick	367	219/148/0	19.30(1.91)	Yes	343	205/138	19.15(1.26)
Wiggins	259	101/157/1	20.85(2.04)	Yes	244	93/151	20.80(1.93)
Total	7,373	2,147/5,175/51	20.77(2.90)		6,404	1,841/4,563	20.38(1.85)

Note: This table contains demographic information for each individual lab in the RRR.
[a]Labs were not considered for the primary analyses if they had less than 100 participants in each condition in the final sample.
[b]Individual participants were not eligible if they (a) did not complete all of the sentence descrambling task items, (b) were not currently a student, (c) did not complete all of the ratings of Ronald, (d) did not complete ratings of all behaviors, (e) were less than 18 years old or older than 25 years old, (f) did not provide gender information, or (g) if there was any "other" information recorded by the experimenters that would exclude them from analyses (e.g., participants did not follow instructions).

Table 2. *List of tasks in combined SW RRR and MAA RRR*

Task	*Description*	*RRR*
Demographics and informed consent	*Provided their age, sex and major and written informed consent*	*[Both]*
Scrambled sentence (hostility priming) (Srull and Wyer, 1979, Exp. 1)	*Mark for 30 groups of 4 words the 3 words that make a complete sentence (e.g., child the question watch). The correct solution was either 80% hostile OR 20% hostile*	*SW*
Vignette (Srull and Wyer, 1979, Exp. 1)	*Read short story about a man named Ronald who behaved in manner that could be seen as hostile (e.g. told a beggar to find a job)*	*SW*
Judgement Ronald (Srull and Wyer, 1979, Exp. 1)	*Judge man from Vignette on 12 characteristics (e.g., Unfriendly)*	*SW*
Judgement Situations (Srull and Wyer, 1979, Exp. 1)	*Judge 15 situations on hostility (e.g., Refusing to let a salesperson into their house)*	*SW*
Abstract Reasoning (Chabris et al., 2018)	*Solve the 10-item version of non-verbal intelligence task*	*[Filler]*
Recall 10 commandments or 10 books (moral reminder)	*Recall the 10 commandments. OR Recall 10 books from high school*	*MAA*
Matrix (cheating opportunity) (Mazar et al., 2008, Exp 1)	*In each of the 20 matrices, find the numbers that add up exactly to 10 (e.g., 3.18 and 6.82). Tear out blank page OR Tear out matrix page*	*MAA*

Collection slip (Mazar et al., 2008, Exp 1)	*List how many matrices solved*	*MAA*
Alternative Uses Test (Guilford, 1967)	*List as many possible uses of a paper clip*	*[Filler]*
Religiousness	*Report religiousness. Specifically, participants were asked to rate, on a scale from 1 (not at all) to 5 (completely), (1) How religious are you? (2) To what extent do you believe in a God? (3) To what extent do you believe in a punishing God?*	*[Preregistered, exploratory moderator of MAA]*
Fatigue (POMS; McNair et al., 1971) and sleep	*Report fatigue and hours of sleep in last night*	*[Exploratory moderator of MAA]*
Time estimation	*Estimate time taken in timed tasks of this battery*	*[Exploratory moderator of MAA]*
HEXACO (Ashton & Lee, 2009)	*Complete 60-item personality scale*	*[Exploratory moderator of MAA]*

Note. This table lists the order of all of the tasks included in the combined Srull and Wyer (1979; SW) Registered Replication Report (RRR) and Mazar, Amir, and Ariely (2008; MAA) RRR.
[a] All between-subjects conditions were counterbalanced

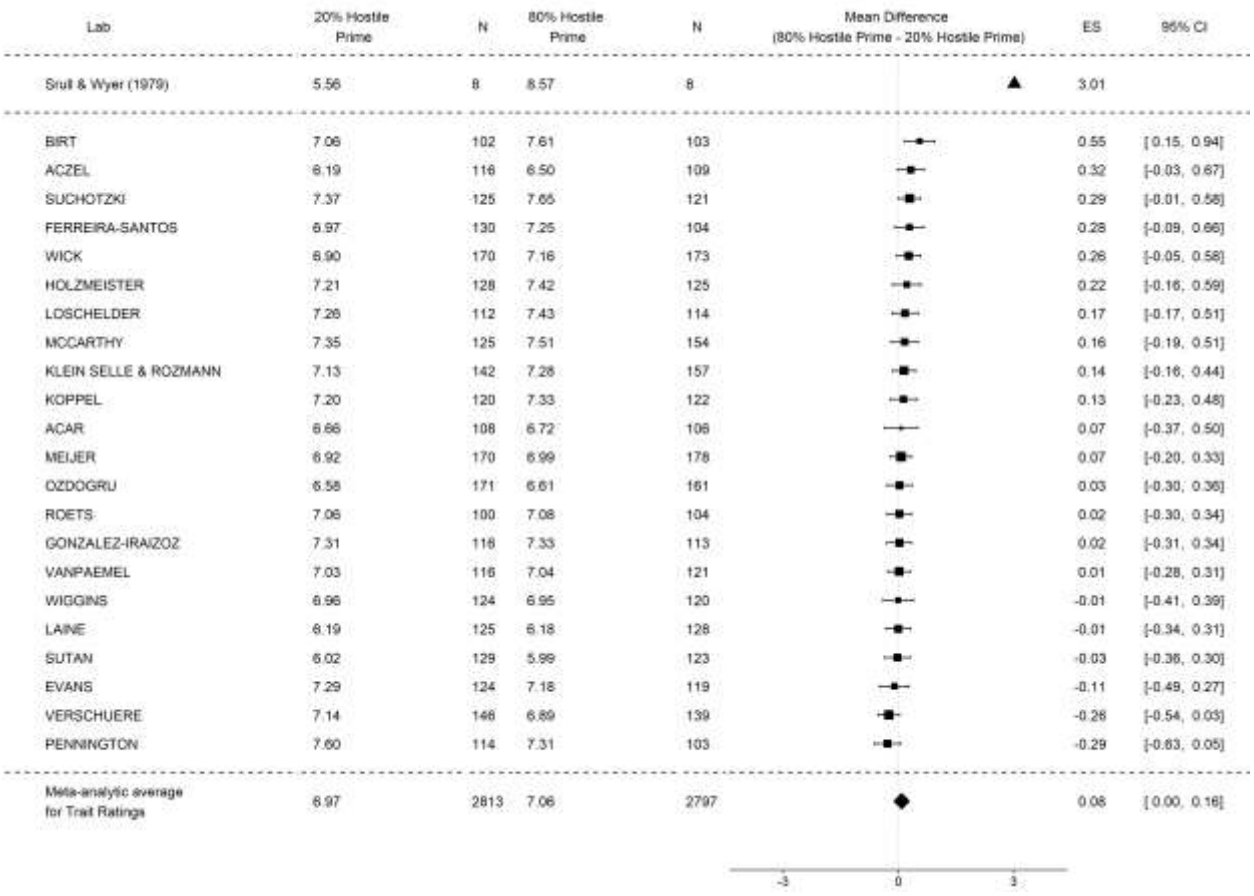

Figure 1. Forest plot of the ratings of "hostile perceptions" of Ronald for the 22 labs included in the primary analyses. The effect size is a mean difference and the error bars represent 95% confidence intervals. The top point represents the estimated effect from Srull and Wyer (1979 [data are no longer available for that effect, and we could not compute confidence intervals from the available information]). The average "hostile perception" for each condition is the unweighted mean of the individual sample means.

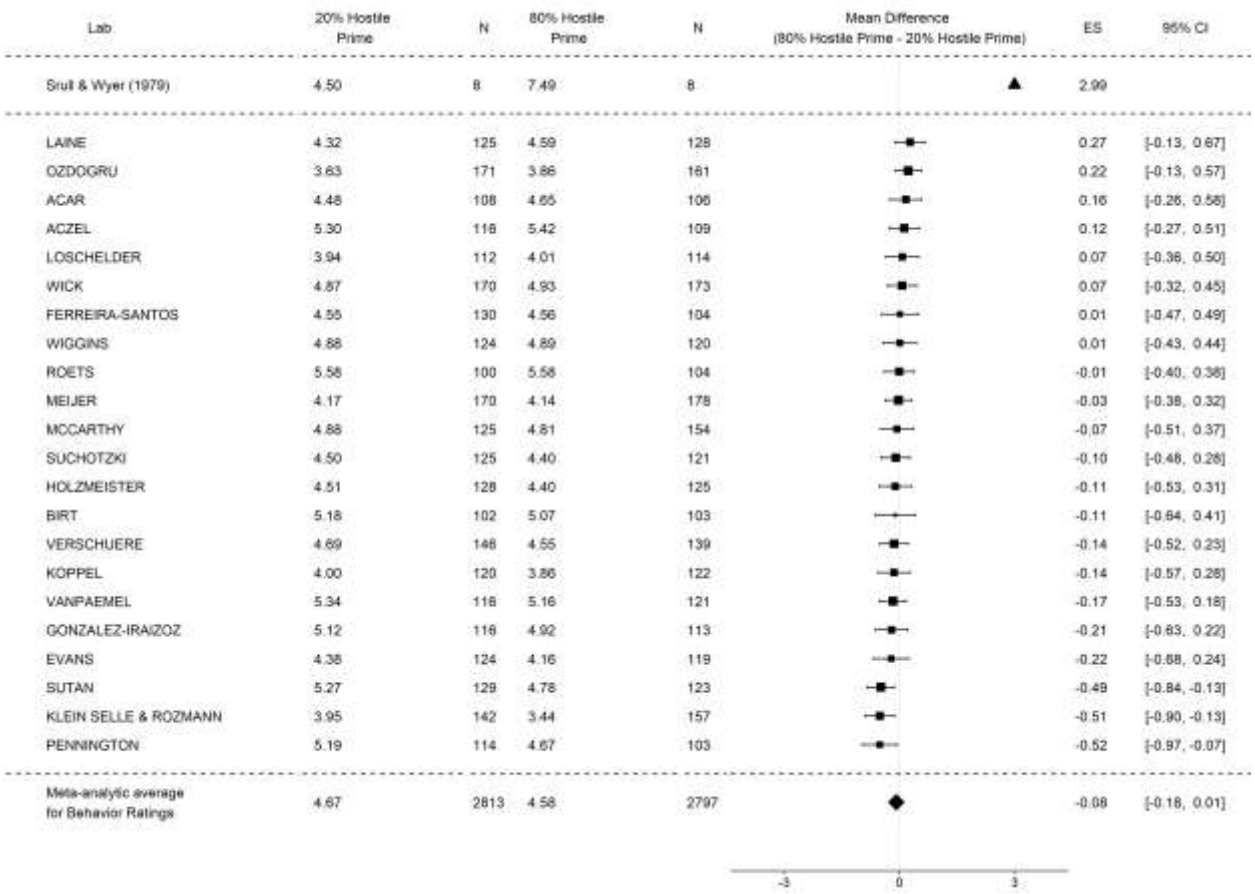

Figure 2. Forest plot of the ratings of hostility for the 5 ambiguously aggressive behaviors for the 22 labs included in the primary analyses. The effect size is a mean difference and the error bars represent 95% confidence intervals. The top point represents the estimated effect from Srull and Wyer (1979 [data are no longer available for that effect, and we could not compute confidence intervals from the available information]). The average rating of hostility for each condition is the unweighted mean of the individual sample means.

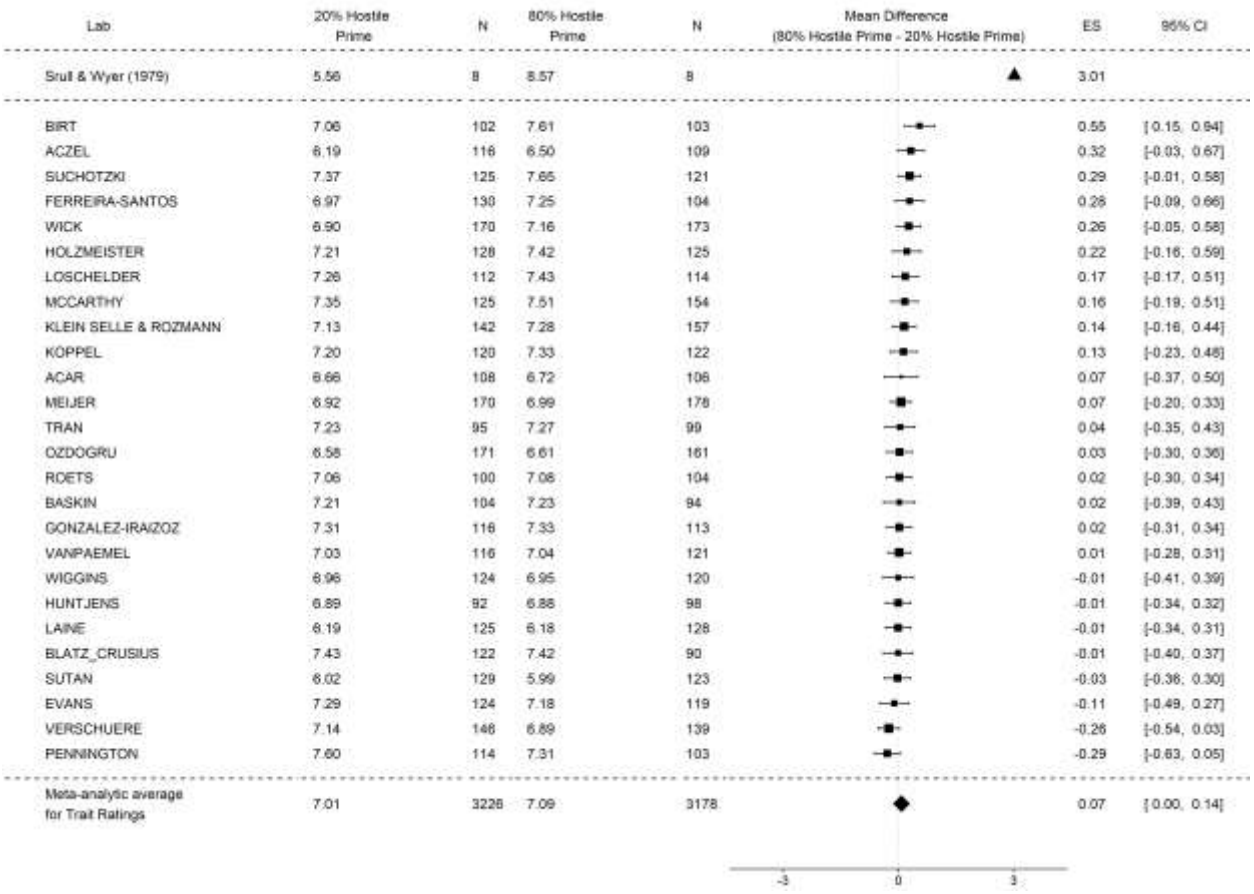

Figure 3. Forest plot of the ratings of "hostile perceptions" of Ronald for the 26 labs included in the ancillary analyses. The effect size is a mean difference and the error bars represent 95% confidence intervals. The top point represents the estimated effect from Srull and Wyer (1979 [data are no longer available for that effect, and we could not compute confidence intervals from the available information]). The average "hostile perception" for each condition is the unweighted mean of the individual sample means.

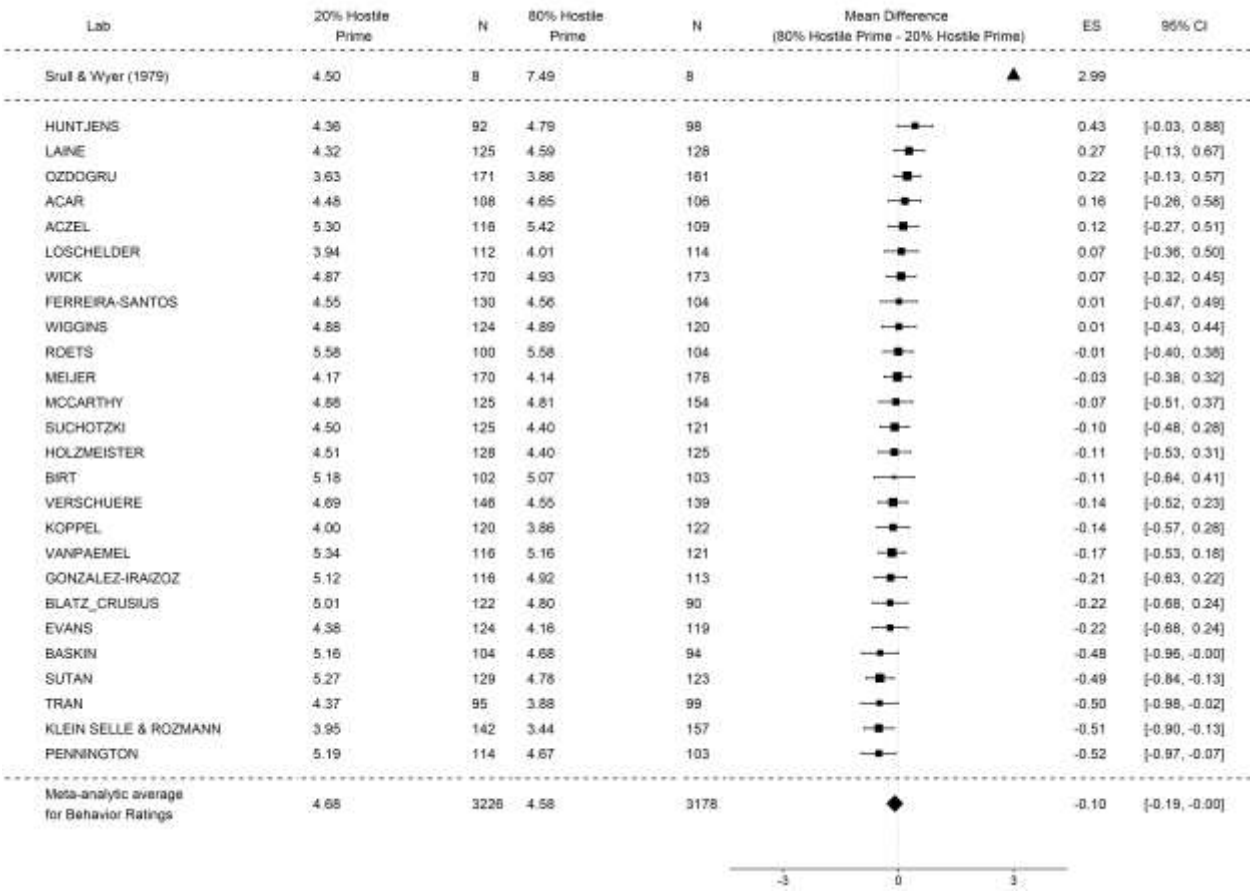

Figure 4. Forest plot of the ratings of hostility for the 5 ambiguously aggressive behaviors for the 26 labs included in the ancillary analyses. The effect size is a mean difference and the error bars represent 95% confidence intervals. The top point represents the estimated effect from Srull and Wyer (1979 [data are no longer available for that effect, and we could not compute confidence intervals from the available information]). The average rating of hostility for each condition is the unweighted mean of the individual sample means.