Incidental Attitude Formation via the Surveillance Task: A Registered Replication Report of Olson and Fazio (2001)

Supplementary Online Materials – Reviewed

Additional details of methods

Sample Size and Characteristics

Table S1 below details the sample size and sample characteristics at each site and percent of exclusions for each of the contingency awareness/recollective memory exclusion criteria. We initially planned that each lab would collect data from a minimum of 100 participants and a maximum of 150 participants based on their local resources. Three labs collected data from more than 150 participants. One lab collected fewer than 100 participants. This was the lab of one of the original authors (Olson). Given that we wanted to offer this lab the opportunity to fully participate in this replication effort, we updated our preregistration with an extended deadline for data collection at this site and specified that all data from all sites would be included regardless of sample size (see osf.io/uyng7 for the addition to the preregistration). This choice was deemed compatible with our meta-analytic approach.

Treatment of Missing Data

Our (preregistered) data processing code excludes participants with missing, partial or incomplete data. However, this exclusion was not explicated in the written preregistration.

Pretesting of Conditioned Stimuli

For the conditioned stimuli, the original authors recommended that we not use the CSs from their original (2001) study because these items may be relatively familiar to modern samples (see Jones et al., 2009). Instead they advised us to select stimuli that would be relatively novel and neutral to the sample population. Based on this recommendation we generated a set of sixty Pokémon characters. We pretested these characters along two dimensions (valence and familiarity) using a separate sample of 155 participants on the Prolific Academic website (https://prolific.ac) (see osf.io/4ecx5). On the basis of this pretest we then selected those twenty characters that were rated as most neutral and least familiar. Participating labs were instructed to further pretest these twenty characters onsite in order to identify the nine characters that are most neutral and least familiar to participants at that specific lab. The two characters that (a) were most neutral and least familiar, and (b) which differed least in valence and familiarity served as CSs (see osf.io/a3qj9 for the results of the pretest conducted at each lab). One lab (Bar-Anan) was unable to carry out such a pretest and therefore used the nine characters derived from the online initial pretest.

Awareness Exclusion Criteria

Primary criterion (Olson & Fazio, 2001). A score was computed following the original authors' recommendations to closely replicate their original study. This score was based on participants' open-ended responses to the original Olson and Fazio's post-experiment question 1 (*Think back to the very first part of the experiment. Did you notice anything out of the ordinary in the way the words and pictures were presented during the surveillance tasks?*) and post-experiment question 2 (*Did you notice anything systematic about how particular words and images appeared together during the surveillance tasks?*). Two independent raters, who were blinded to one another's ratings, evaluated responses to these two questions, and treated responses on both questions as one (compound) text response (see osf.io/2dm6u for the exact coding instructions provided to the data collection sites). Specifically, they scored participants as being 'aware' if their responses to either of these two questions made correct reference to *both* of the CS-US pairings. In other words, they were scored as 'aware' if they wrote that CS_{pos} (either its name or a description of its appearance) appeared during the task together with positively valenced words/images *and* that CS_{neg} (its name or a description of its appearance) appeared during the task together with negative words/images. If they failed to meet this criterion for any reason then they were scored as 'unaware'. This included (a) identifying only one of the two CS-US pairings, (b) identifying the CS-US pairings incorrectly (i.e., reversed), (c) identifying that the two CS were paired with US stimuli but not specifying which was paired with which, or (d) not identifying CS-US pairings at all. Scores were then compared between raters to assign each participant a single score. Participants were only scored as 'aware' if they were scored by both raters as being 'aware'.

Secondary criteria. We considered that the original authors' criterion may have scored individuals who were actually aware of/remembered the contingencies as 'unaware'. Therefore we preregistered three additional secondary exclusion criteria that allowed us to examine if evidence for EC effects in this task were robust to or depended on the specific way in which contingency awareness/recollective memory was measured. The exact instructions provided to the data collection sites for the 'Olson and Fazio (2001) modified' criteria can be found at osf.io/2dm6u. Data processing for the 'Bar-Anan et al. (2010)' and 'Bar-Anan et al. (2010) modified' criteria required no hand scoring and were performed algorithmically (see osf.io/k9nrf for R script).

Secondary criterion 1 (Olson & Fazio 2001 modified). This criterion was identical to the Olson and Fazio (2001) criterion with one modification: participants were scored as 'aware' if their responses to the two questions referred to *any form of systematic pairing* between the CS and US stimuli, regardless of whether specific pairings were described. Specifically, participants were coded as "aware" if they (a) identified only one of the two CS-US pairings, (b) identified that the two CS were paired with US stimuli but not specifying the specific way in which the CSs and USs were paired. Participants were coded as "unaware" only if their answer did not contain any mention of a systematic pairing between CSs and USs. In cases of disagreement between the two raters, the participant's responses were scored by a third rater. The participant was scored as 'aware' or 'unaware' based on the majority judgment.

Secondary criterion 2 (Bar-Anan et al., 2010). This criterion was computed based on Bar-Anan et al.'s (2010) criterion. Here participants were asked: *For some participants, during the first task, there was one cartoon creature that always appeared with positive images and words, and one that always appeared with negative images and words. Do you think it happened in your case?* (Question 1 from the Bar-Anan et al. protocol). They were scored as 'aware' if they responded "Yes, that happened in my task" and as unaware if they chose "No, I did not notice if that happened in my task".

Secondary criterion 3 (Bar-Anan et al., 2010 modified). This criterion was identical to the Bar-Anan et al., (2010) criterion with the addition that participants had to correctly identify (on post-experiment questions 2 and 3 from the Bar-Anan et al. protocol) the valence of the USs with which each of the two CSs appeared. Specifically, post-experiment questions 2 and 3 from the Bar-Anan et al. protocol presented participants with images of the two CSs and asked them the following: *During the first task, which of the two characters was consistently presented with [positive/negative] images and words?* Response options were, for example, "BERGMITE (certainly)", "BERGMITE (probably)", "BERGMITE (guess)", "PALPITOAD (guess)", "PALPITOAD (probably)", "PALPITOAD (certainly)"). Note that the specific Pokémon exemplars used in the questions depended on those used at each laboratory. Participants were scored as 'aware' if they identified the correct CS that was paired with the US *and* used either the "probably" or "certainly" response options when doing so (i.e., not the "guess" option). All other participants were scored as 'unaware'.

Deviations from the Preregistration

In order to maximize evidential value and transparency, we document all divergences from the preregistration/Stage 1 accepted manuscript (see osf.io/kzchq/) to Stage 2 manuscript below.

Change in Terminology from 'Confirmatory'/'Exploratory' to 'Primary'/'Secondary' Analyses

After writing the Stage 2 manuscript and soliciting comments from the co-authors, there was consensus that the terminology of 'confirmatory' vs. 'exploratory' analyses was confusing given that all analyses were preregistered (both descriptions and the code implementing them). However, we were also acutely aware of the potential pitfalls of relabeling these analyses given the Registered Report format. We therefore sought advice from Christ Chambers, creator of the Registered Report format and editor for a large number of RR articles to date, about the relative benefits and costs of changing vs. not changing this terminology. His expert opinion was that the term 'exploratory' should not be employed within a preregistered analysis. As such, we have changed the Stage 2 manuscript to refer to 'primary' analyses (i.e., those that most directly replicate the original Fazio & Olson, 2001 study) versus 'secondary' analyses (i.e., those that test the robustness of the EC effect to other exclusion criteria). We felt that this modification to the Stage 1 accepted manuscript was justified on the basis of improving clarity and readability. This change, along with reference to this document, is now footnoted in the manuscript.

Interpretation of the Results

When we came to the interpretation of the results based on our preregistered criteria, we realised that there was an incongruence between the analyses we had pre-registered and interpretations of these analyses that we had pre-registered. At this point we realized that a deviation from preregistration of some form was unavoidable. Given that the interpretation of results is central to the article, we therefore describe here a) what the Stage 1 Accepted manuscript stated, b) the incompatibility between our stated plans for analysis and interpretation, c) our priorities and goals when considering how to resolve this issue, and d) the strategy we adopted to do so.

Plan as stated in Stage 1 Accepted manuscript. Note that the below quotes are verbatim, retaining the original language of 'confirmatory' vs. 'exploratory' analyses and hypotheses. However, we list them under the headings of what the article now refers to as 'primary' vs. 'secondary' analyses (see above).

Primary analyses and hypotheses. *"To determine if EC effects emerge in the absence of contingency awareness/recollective memory, according to the original authors criteria, we will compute the EC effect size (Hedges' g) from the mean and standard deviation of the self-reported preference score in the 'unaware' group.*

Thereafter we will meta-analyze these effect sizes in a meta-analysis using a random-effects model, using an alpha value of 0.05. Although all participating labs will use similar materials, differences may be introduced by the translation of materials, selection of stimuli, or characteristics of the samples. In order to account for this within the analyses, we will employ random effects meta-analysis models (specifically, using the Restricted Maximum Likelihood method)." (p.14)

R code implementing this exact meta-analytic model was preregistered with the Stage 1 Accepted Manuscript and was not subsequently changed (see osf.io/3hjpf for all preregistered code, and osf.io/qga5j/ for all finalized code), i.e.,

```
fitted_model <-
    rma(yi   = hedges_g,
        sei  = hedges_g_se,
        data = data_effect_sizes,
        slab = data_collection_site)
```

"Based on the above analyses, these findings [replicate/do not replicate] the original authors findings." (p.14)

Secondary analyses and hypotheses*. "Three different groups will be created (i.e., those based on the modification to the original authors' criteria, those based on the original Bar-Anan et al., criteria, and those based on the modified Bar-Anan et al. criteria). For each group (in each lab) we will compute the EC effect size (Hedges' g) from the mean and standard deviation of the self-reported preference score. Thereafter we will meta-analyze these effect sizes in three independent meta-analyses using a random-effects model."* (pp.14-15)

R code to implement these analyses was also preregistered. Specifically, the same R code used for the primary analysis above was employed, changing only the data being passed to the same function.

"Finally, to investigate if the effect sizes computed based on the four awareness/recollective memory criteria differ from one another, we used a multilevel meta-analysis with the type of criteria as a moderator, adding a random intercept for laboratory to account for the statistical dependency between effect sizes coming from related samples." (p.16)

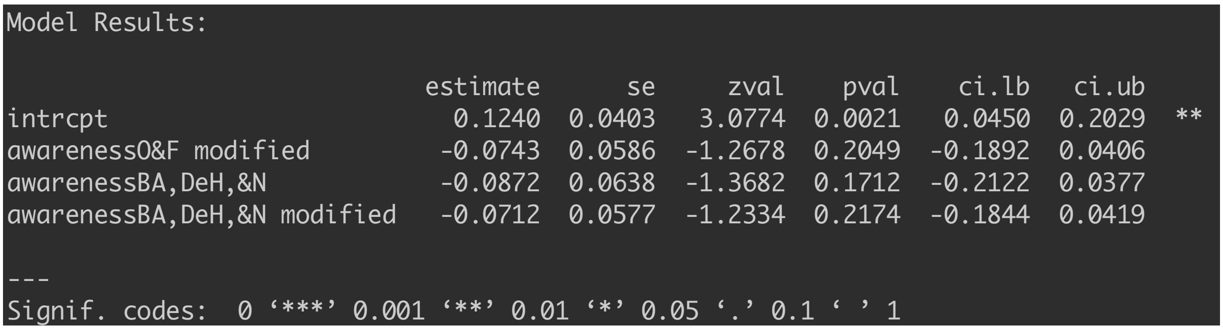
R code implementing this exact meta-analytic model was preregistered with the Stage 1 Accepted Manuscript and was not subsequently changed, i.e.,

```
fitted_model <-
    rma.mv(yi     = hedges_g,
           V      = hedges_g_se^2,
           mods   = ~ awareness,
           random = ~ 1 | data_collection_site,
           data   = data_effect_sizes,
           slab   = data_collection_site)
```

"There are three outcomes that we have a priori hypotheses for. The first is a situation where the multilevel meta-analysis returns a significant overall EC effect, but no significant effect for the type of criteria. In this case, we will conclude that EC effects do emerge in the surveillance task and do not depend on the specific way in which contingency awareness/recollective memory is measured. The second is where we find no evidence for an overall EC effect and the type of criteria also fails to moderate the size of EC. In this case, we will conclude that EC effects do not emerge in the surveillance task. The third is where we find a significant effect of type of criteria in the multilevel meta-analysis and the individual univariate meta-analysis reveal significant evidence for EC with the original authors' criteria but with none of the other three criteria. In this case, we will conclude that EC effects in the surveillance task strongly depend on the way that the original authors chose to assess contingency awareness/recollective memory." (p.17)

Incompatibilities detected in the preparation of the Stage 2 manuscript. Critically, the previously quoted paragraph, which describes how results of the individual meta-analyses and the multilevel moderator meta-analysis will be integrated, refers to results that our pre-registered model does not produce. Specifically, a multilevel moderator meta-analysis model does not produce an estimate of an "overall EC effect", but rather four separate estimates of the EC effect using each of the four exclusion criteria.

To recount the logic and action of the method here, a standard univariate meta-analytic model is effectively an intercept-only model (in terms of its 'fixed' effects). When extended to a moderator meta-analysis model, an additional fixed effect is added to the model. The standard coding strategy to implement this is to treat one of the levels of the moderator (i.e., the exclusion criteria) as the intercept (i.e., as a reference category), with the other levels of the moderator estimated as main effects. Our preregistered code defined the Olson and Fazio (2001) criterion as the intercept. The meta-effect size for any of the three criteria is therefore calculated as intercept + main effect for that criterion. The key point to be appreciate here is that no ‘overall’ effect is estimated: estimates are made for each of the exclusion criteria. This is perhaps best understood by seeing the output itself:



```
Model Results:

                             estimate      se     zval    pval    ci.lb   ci.ub
intrcpt                        0.1240  0.0403   3.0774  0.0021   0.0450  0.2029  **
awarenessO&F modified         -0.0743  0.0586  -1.2678  0.2049  -0.1892  0.0406
awarenessBA,DeH,&N            -0.0872  0.0638  -1.3682  0.1712  -0.2122  0.0377
awarenessBA,DeH,&N modified   -0.0712  0.0577  -1.2334  0.2174  -0.1844  0.0419

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

How did this error arise? While we cannot fully account for this oversight, we think it is likely that this error in planned interpretation arose through an incorrect analogy with ANOVA results when writing this section of text. Specifically, “moderation” in the context of ANOVA is typically quantified through separate “main” and “interaction” effects. Output from such tests therefore typically provides the researcher with an understanding of what could be called the ‘overall’ effect and also its moderation by another variable (i.e., in the interaction effect).

What are its implications? Two of the three combinations of outcomes from the multilevel and univariate meta-analyses referred to situations in which a “significant overall EC effect”, when in fact this meta-analysis model does not quantify any such effect. Specifically, outcome combinations 1 (significant overall EC effect & non-significant moderation by criterion) and 2 (non-significant overall EC effect & non-significant moderation by criterion; see above for full quotes and p.17 of Stage 1 Accepted manuscript). These criteria could therefore not be fulfilled when interpreting the results. The third outcome combination specified (significant moderation by criterion & significant evidence for EC effect using Olson & Fazio 2001 criterion and non-significant EC effect using the other three criteria) was not met by the results (i.e., no moderation by criterion was found).

What was our solution? Our solution was to stick to our preregistered plan, while acknowledging its limitations. First, given the precision of our planned analyses both in terms of their written description and their code implementation, we did not change these in any way. Second, we acknowledge that the written plan on how to integrate the interpretation of the results of these tests did not correctly correspond to the output of these tests. Third, despite this mismatch, we considered it appropriate to stick to our preregistered plan: the main article therefore does not conclude in favour of any of outcome combinations 1, 2, or 3 for the secondary analyses. Instead, it notes that there is thus great uncertainty regarding whether EC effects differ between the four criteria (these points can be found in the discussion section and will not be reproduced here). It is important to note that the written plan to interpret the results of the primary analyses (i.e., whether the original effect was replicated) was unaffected here.

Data Collection Stopping Rule

Due to unforeseen delays, one site was unable to collect data from the specified planned number of participants (100 to 150 per site) within the informally agreed upon timeframe. We provided this lab will additional time insofar as was possible. However, we realized that no maximum timeframe was specified in our preregistration. In order to resolve the situation, we made an updated preregistration that modified our data collection stopping rule (see osf.io/uyng7). This updated preregistration (made on 2020-02-11) is discussed in the manuscript. It specified that we would instead use all data collected from all sites, even those who had not met the originally planned sample sizes, and set a hard deadline for data collection after which any and all data from each site would be used (2020-02-19). This also accommodated sites that collected data from more participants than planned in our preregistration. This modification was deemed to be consistent with our meta analytic approach within the preregistered analyses (i.e., even small samples sizes make meaningful contributes as the estimation of the meta effect size, as the uncertainty around all effect sizes is quantified within the meta-analysis models). This decision was driven in large part by the fact that this lab was that of one of the original authors, who we felt it was therefore particularly important to include in the replication.

Method of Calculating Confidence Intervals

The preregistered implementation of the analyses employed a bootstrapping method to calculate effect sizes at each site prior to meta-analyses. However, due to the change in the data collection stopping rule (see above) one site collected a far smaller than predicted number of participants ($n = 21$). Heterogeneity metrics (e.g., I^2 and H^2) were observed to computationally unstable when re-running the analysis script. For the sake of computational reproducibility, we therefore exchanged the bootstrapping method for the arithmetic method throughout. Inspection of the effect sizes and CIs suggested the impact of this decision on the meta effect size estimates and its confidence intervals was less than Hedge's $g = 0.01$. We also note that the written description included in the Stage 1 Accepted manuscript implied that the arithmetic method would be employed (i.e., "we will compute the EC effect size (Hedges' *g*) from the mean and standard deviation", p.14).

z and *p* Values for 'Aware' Participants

Due to an oversight, no method to calculate *z* or *p* values was specified or implemented in our written preregistration or preregistered code. The preregistered implementation of the moderator meta-analysis models return values for the difference in effect size between the two subsets ('aware' vs. 'unaware') and the *p* and *z* values for this difference, but not values for each subset. While the preregistered models are fit for their primary purpose (i.e., testing moderation by awareness), our Stage 1 accepted manuscript stated that we would also report *z* and *p* values for each subset. In order to employ the identical method to how these values were calculated for the 'unaware' subset, we therefore fitted (non-moderator) meta-analyses to just the aware subset of participants. The only results reported from these models were the effect sizes, 95% CIs, *z* and *p* values.

Description of the exclusion criteria

After writing the Stage 2 manuscript and soliciting comments from co-authors, there was consensus that the description of the four exclusion criteria was unclear and confusing. We therefore elected to rewrite this section (see pp. 11-13 in the manuscript). Following the editor request, we moved most of the description of the four exclusion criteria to the SOM-R (see pp. 2-4 above). Importantly, it is only the *description in the manuscript* of what these criteria consisted of and how they were applied that changed. Their implementation did not change between preregistration/Stage 1 acceptance and the Stage 2 manuscript. In fact, the revised descriptions of the criteria in this section are more closely aligned with the actual preregistered protocol and instructions distributed to the sites than the descriptions in the Stage 1 accepted manuscript. We therefore felt this this modification to the Stage 1 accepted manuscript was justified on the basis of improving clarity and readability.

Non-Preregistered Analyses

All non-preregistered analyses are clearly marked in both the code implementation and the manuscript.

Additional Details on Results

Comparison of 'Contingency-Aware' vs. 'Unaware' Participants

The initial set of analyses in our paper always excluded 'contingency-aware' participants. Yet one could also examine whether awareness/recollective memory moderate the size of EC effects. With this in mind, we divided participants into two groups ('aware' and 'unaware') using the four aforementioned criteria, and then carried out an additional set of secondary analyses that compared EC effects between these two groups using a multilevel moderator meta-analysis model.

Non-Preregistered Analysis: Moderator Meta-Analysis

After data collection and analysis, a co-author pointed out that the assumption of independence was violated within our moderator meta-analysis, given that two of the exclusion criteria (Olson & Fazio, 2001 modified and Bar-Anan et al., 2010 modified) are derivatives of the other two (Olson & Fazio, 2001 and Bar-Anan et al., 2010). In order to ensure that this violation did not influence conclusions, we fitted one additional model. This was identical to the moderator meta-analysis model with one exception: instead of treating the criteria as one variable with four levels, it treated them as two: criterion 'family' (i.e., Olson & Fazio type vs. Bar-Anan et al. type) and 'strictness' (i.e., one of the two within each family was stricter than the other). These two variables and their interaction were included as moderators in the meta-analysis model. Consistent with the results of the preregistered moderator meta-analysis model, no evidence of moderation was found either overall, $Q(3)$ =2.76, $p = .430$, or for the change in meta effect sizes for family, strictness, or their interaction, all *p*s ≥ .205.

Table S1. Sample size, sample characteristics, and percent of exclusions for each of the contingency awareness/recollective memory exclusion criteria, as a function of data-collection site.

			Age		Gender				Percent excluded				
Site	*n* manual exclusions	*n* for analysis	Mean	SD	Female	Male	Other identity	Did not answer	Surveillance task performance	Olson & Fazio (2001)	Olson & Fazio (2001) modified	Bar-Anan et al. (2010)	Bar-Anan et al. (2010) modified
Balas	6	100	26.5	4.7	57	43	0	0	3.0	2.1	19.6	41.2	16.5
Mierop	1	99	21.7	4.2	66	33	0	0	2.0	8.2	17.5	43.3	21.7
Gast	0	120	23.6	7.2	91	26	1	2	2.5	6.0	26.4	49.4	24.7
Gawronski	0	155	18.9	1.1	113	41	1	0	2.6	7.2	74.1	51.2	30.2
Hütter	2	148	22.7	6.2	109	39	0	0	1.4	18.4	41.6	57.3	43.6
Kurdi	0	151	19.3	1.3	120	31	0	0	1.3	8.0	21.4	39.4	21.4
Moran	1	99	20.0	3.2	75	24	0	0	1.0	2.0	28.6	46.9	27.6
Olson	0	21	20.0	0.0	10	11	0	0	0.0	9.5	28.6	42.9	33.3
Douglas	0	148	18.6	0.8	98	50	0	0	2.0	6.9	19.9	58.2	35.6
Stahl	0	100	21.7	5.1	80	20	0	0	3.0	13.4	32.0	54.6	35.1
Unkelbach	0	142	23.6	7.0	82	57	1	2	1.4	10.0	36.3	51.2	29.9
Vadillo	0	195	19.9	3.0	166	25	3	1	1.5	1.0	15.0	39.3	12.9

Note. Each lab is identified by the last name of the corresponding author. *n* manual exclusions: exclusions made manually before the analysis due to incomplete data file (1 case at Moran's site, 2 cases at Hütter's site), technical problems (4 cases at Balas's site), unusual participant behaviour (1 case at Balas's site), participant eligibility (1 case at Balas's site), and data recoding issues (1 case at Mierop's site). *n* for analysis: represents the sample size after the manual exclusions. Age and gender are characteristics are calculated from the sample for analysis after manual exclusions. Percent excluded surveillance task performance: percent of exclusions based on the number of errors made during the surveillance task (percentage accuracy < mean – 3 SD per site). Percent excluded for Olson & Fazio (2001), Olson & Fazio (2001) modified, Bar-Anan et al. (2010), and Bar-Anan et al. (2010) modified represent the percent of the sample excluded *after* surveillance task exclusions had been excluded. These mirror the way these exclusions have been reported in the manuscript.

References

Bar-Anan, Y., De Houwer, J., & Nosek, B. A. (2010). Evaluative conditioning and conscious knowledge of contingencies: A correlational investigation with large samples. *The Quarterly Journal of Experimental Psychology, 63*(12), 2313-2335.

Jones, C. R., Fazio, R. H., & Olson, M. A. (2009). Implicit misattribution as a mechanism underlying evaluative conditioning. *Journal of Personality and Social Psychology, 96*(5), 933-948.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12*(5), 413-417.