Breakthrough or One-Hit Wonder?

Three Attempts to Replicate Single-Exposure Musical Conditioning Effects on Choice Behavior (Gorn, 1982)

Ivar Vermeulen,[1] Anika Batenburg,[1] Camiel J. Beukeboom,[1] and Tim Smits[2]

[1]Faculty of Social Sciences, Communication Science, VU University Amsterdam, The Netherlands,
[2]Faculty of Social Sciences, Institute for Media Studies, Katholieke Universiteit Leuven, Belgium

Abstract. Three studies replicated a classroom experiment on single-exposure musical conditioning of consumer choice (Gorn, 1982), testing whether simultaneous exposure to liked (vs. disliked) music and a pen image induced preferences for the shown (vs. a different) pen. Experiments 1 and 2 employed the original music, Experiment 3 used contemporary music. Experiments 2 and 3 employed hypothesis-blind experimenters. All studies incorporated post-experimental inquiries exploring demand artifacts. Experiments 1 and 2 (original music; $N = 158$, $N = 190$) showed no evidence for musical conditioning, and were qualified (conclusive) replication failures. Experiment 3 (contemporary music; $N = 91$) reproduced original effects, but with significantly smaller effect size. Moreover, it had limited power and showed extreme scores in one experimental group. Aggregated, the three studies produced a null effect. Exploration of demand artifacts suggests they are unlikely to have produced the original results.

Keywords: music in advertising, musical conditioning, demand characteristics, direct replication, Gorn

This paper focuses on replicating the first experiment in Gerald Gorn's article "The effects of music in advertising on choice behavior: A classical conditioning approach," published in the *Journal of Marketing*, 1982. The original experiment's findings are taken as evidence that music can unobtrusively, and through single exposure, condition consumer choice behavior. The study has an almost iconic status and is impressively proliferated through the literature. Google Scholar reports 662 citations, Web of Knowledge 243 (October 30, 2013). Moreover, the study appears in nearly every student textbook on persuasion and consumer psychology (e.g., Cialdini, 2001; Fennis & Stroebe, 2010; Peck & Childers, 2008; Saad, 2007). Several replications of the original study failed, but, as will be described shortly, none exactly followed the original procedures. A direct replication is still lacking.

The original experiment (Gorn, 1982, Experiment 1) involved a 2 (Pen color: light blue vs. beige) × 2 (Music: liked vs. disliked) between-subjects design, conducted among 244 undergraduate students in a management course. Participants were asked, during class time, to evaluate a piece of music that an advertising agency considered for a pen commercial. Depending on the experimental condition, they were then exposed to a picture of a light blue or beige pen on a big screen while either "liked" or "disliked" music (an excerpt from the movie Grease vs. classical Indian music) played for 1 minute. To thank participants, they were offered a light blue or a beige pen (one of which was previously "advertised" on screen). Upon leaving the classroom they could choose a pen from one of two boxes, with question sheet drop-off boxes next to them.

Results showed that 79% of participants in the "liked" music conditions chose the pen in the color displayed on screen. Only 30% of participants in the "disliked" music conditions chose the displayed pen. Furthermore, when asked afterwards for their reasons to choose a particular pen color, only 2.5% of participants mentioned the music. These findings suggest that simple, fairly unobtrusive cues like music can influence consumer choice behavior following single exposure.

Replicating Gorn's experiment is important, not only because of its impact on the persuasion literature, but also because its rather unconventional procedure has repeatedly been criticized (e.g., Allen & Madden, 1985; Kellaris & Cox, 1989). Replicating the study constitutes a challenge, as it involves stimuli susceptible to cultural trends and changes (e.g., musical preferences).

Existing Evidence

The original study (Gorn, 1982) produced a rather strong (Cohen, 1988) effect size: $\varphi = .49$. Some found this noteworthy because it does not seem to employ a particularly powerful conditioning procedure (e.g., Bierley, McSweeney, & Vannieuwkerk, 1985). Participants were exposed to stimuli only once instead of repeatedly;

although conditioning effects have been shown after single trials (Stuart, Shimp, & Engle, 1987, Experiment 1), such effects usually require very strong stimuli, like nauseating drugs or intense shocks. Moreover, in Pavlovian conditioning, strongest results are usually reported when conditioned stimuli (pens) are presented before unconditioned stimuli (music) rather than simultaneously.

Some scholars suggested that the strong effects found in the original study originated in demand artifacts (Allen & Madden, 1985; Kellaris & Cox, 1989). Participants' awareness of the study's purpose may have elicited behavior congruent to inferred expectations. Gorn inferred participants did not know the study's purpose, as only five out of 205 mentioned music as a reason for pen choice. However, Allen and Madden (1985) commented that the original post-experimental inquiry lacked detail and rigor.

To elucidate these issues, several scholars conducted replication studies. The literature contains two studies demonstrating results congruent to the original (Bierley et al., 1985; Groenland & Schoormans, 1994), and three that failed to show significant findings (Allen & Madden, 1985; Kellaris & Cox, 1989, Experiments 1 and 3). However, the experimental procedures of all replications (both successful and unsuccessful) included fundamental modifications that could have caused different findings. For example, experiments were conducted in cubicles instead of a classroom, researchers used alternative liked and disliked musical, or even nonmusical, stimuli (e.g., humor segments), participants were offered pens from one rather than two boxes, exposed to one color pen only, or asked to answer questions about the advertised pens' characteristics before pen choice. Therefore, none of these studies can be considered direct replications. For a more extensive review of prior replication attempts, see the preregistered proposal (http://osf.io/z6e8j).

Aforementioned replication attempts failed to show conclusive evidence that attributes the original findings to demand characteristics. Shimp, Hyatt, and Snyder (1991, 1993) conclude, after analyzing the failed replications by Kellaris and Cox (1989), that the original findings (Gorn, 1982) more likely originate in successful conditioning than in demand artifacts. In sum, several replications of the original study were conducted but none were direct, and the aggregated knowledge remains inconclusive.

Study Outline and Power

The current replication attempt follows a three-step approach. Experiment 1 replicates the original procedures using (close to) original materials and (following the original study) a fully informed experimenter team. Experiment 2 employs the same procedures and materials, but uses non-informed experimenters; it also employs more extensive post-experimental questionnaires to explore demand characteristics. Experiment 3 mirrors the basic procedures of Experiment 2, but uses contemporary rather than the 1980s musical stimuli.

Planned sample sizes (250, 240, and 200, respectively) were based on estimated attendance of the classes in which the experiments would take place, and on a priori power analysis using *G*Power* (Erdfelder, Faul, & Buchner, 1996). This analysis showed that, assuming $\alpha = .05$ and inclusion of 80% of participants in the main analyses (cf. the original study), $N = 69$ would suffice to detect the original effect size ($\varphi = .49$; logOR = 2.17, 95% CI = 1.52–2.82)[1] with .95 power. Recently, Simonsohn (2013) suggested samples for replication studies should multiply the original sample by 2.5 (thus, in this case, $N = 610$ per study) to enable reliable detection of the effect size that would have given the original study .33 power (in this case, $\varphi_{33\%} = .11$; logOR = .44). However, such sample sizes (classes of 610 students) are unattainable in the current experimental set-up. For replications of large studies, Simonsohn suggests testing results against a practically or theoretically "small" point null effect size. We found no theoretical footholds to determine such a point null. Based on a simple return-on-investment advertising scenario, we set a practical point null effect size at half the original effect size ($\varphi = .25$; logOR = 1.04). Note however, that this point null is fairly arbitrary, and was set post hoc to enable categorization of replication outcomes (cf. Simonsohn, 2013).

Due to falling student numbers and low class attendance (for Experiment 3), we did not reach planned sample sizes, but effective samples of $N = 158$, $N = 190$, and $N = 91$ instead. Note that these samples still provide > .95 power to detect the original effect. Actual obtained power and sensitivity analyses will be presented with each sample.

Experiment 1

Experiment 1 directly replicated Gorn's original study including (1) the original musical stimuli most likely used, (2) pens in two pretested colors, (3) a balanced 2 (Music: liked vs. disliked) × 2 (Advertised Pen: color 1 vs. color 2) design, (4) original instructions and procedures, and (5) presence of an experimenter team aware of the hypothesis tested.

Participants

Participants were 160 second year BA Communication students, recruited through attendance of a persuasion class taught at a Belgian university. At onset, participants were unaware of the study's purpose, as (1) no informed consent was asked, (2) the experiment was not announced, (3) no prior references to the original study had been made earlier in the participants' curriculum. Students did not receive credits or money for participation.

[1] We will report log odds ratios (logOR) for all χ^2 tests; logOR's are approximately normally distributed, and therefore easy to interpret (logOR = 0 indicates no effect, and logOR is in the center of its CI; Bland & Altman, 2000).

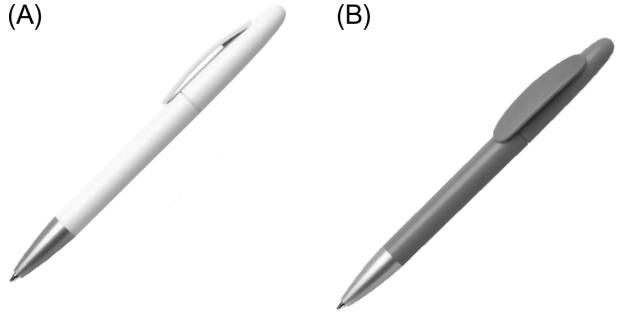

Figure 1. White pen (A), light blue pen (B).

Materials

Pens

In an online pretest, 48 Dutch MA Communication students (M_{age} = 22.77, SD = 3.83; 19 male, 28 female, 1 unknown) judged 13 pens differing only in color. Of the pens scoring similar to the collective mean (which excluded beige), the difference between white and light blue pens was the smallest in all possible pairs (M = 4.31 vs. 4.25; scale 1–7; $F(1, 47)$ = .04, 95% CI_{diff} = −.57 to .70, p = .84, η^2 = .00) – hence these were selected as stimuli. See electronic supplementary materials for details. Pens were offered in two unmarked carton boxes holding 150 pens. For display on the slides, professional pictures from the manufacturer's website were used (see Figure 1).

Music

The original paper left the exact musical stimuli unspecified. Aided by the original author, we selected five songs from the musical Grease (liked music) and five classical Indian songs by Parveen Sultana (disliked music), which we pretested in 47 students (M_{age} = 22.00, SD = 1.98; 21 male, 26 female; 29 Dutch, 18 Belgian; see supplementary materials). Based on this pretest, and avoiding possible lyrical confounds, we selected the Grease song "Summer Nights" (M = 5.37; scale 1–7) as liked music and the Parveen Sultana song "Aaj Kaun Gali Gayo Shyam" (M = 2.56) as disliked music. Both differed significantly in evaluation, $F(1, 46)$ = 126.48, 95% CI_{diff} = 2.31–3.31, $p < .001$, η^2 = .73.

Procedure

The informed experimenter team consisted of one of the authors (teaching the class) and four non-naive assistants.

Participants received an announcement a week prior to the experiment, explaining that the class would be split in two due to scheduling problems; division was approximately 50/50 based on alphabetical order. Halfway each class, students with odd student ID numbers were asked to follow an assistant to a waiting room, while even numbered students stayed. After completing the study, even numbered students left for another waiting room while odd numbered students returned.

While two assistants distributed the music evaluation forms, the experimenter explained – following the original script – that an advertising agency was trying to select music to use in a commercial for a pen produced by one of its clients. Participants would hear some music that was being considered while they would see an image of the pen that the agency was planning to advertise on a PowerPoint slide.

While displaying the pen for one minute, a music excerpt was played over the class' sound system. Afterwards, participants evaluated the music on the form. Subsequently, they were told that they would receive either a white or a blue pen for their help, donated by the manufacturer. The experimenter held up each pen briefly and commented that if they wanted a white one, they should go to the box positioned left of the class room's exit, whereas if they wanted a blue one, they should go to the box positioned on the right[2], and drop off their question form next to the boxes. Participants were invited to line up for the exit, thus exposing them equally to both boxes. After collecting their pens, participants were given a brief questionnaire with four open questions (see below) and were asked to indicate pen choice, age, gender, and the last three digits of their student ID. Afterwards, the class continued. Debriefing took place in the subsequent class.

Measures

Music evaluation was measured using three 5-point Likert scale items (see supplementary materials; α = .93).

Pen choice was measured by (1) assessing at which table music evaluation forms were handed in, (2) an unobtrusive code on the post-choice question forms handed out at the same tables, and (3) participants' self-reported pen color choice on the post-choice form. Each form asked for age, gender, and last three student ID digits, enabling us to link responses. Participants with conflicting pen choice measurements were excluded.

Reasons for pen choice were measured using three open answer boxes. See supplementary materials for details.

Hypothesis awareness – participants' awareness of the hypothesized relationship between music played and pen choice – was assessed in one open question asking participants about the goal of the session they just attended. Answers were coded between 0 and 5 for increasing hypothesis awareness (see supplementary materials for details).

[2] Here, we diverted slightly from the procedure described in the original paper, which stated that the boxes were positioned on the left and right side of the classroom. In response to our concerns that this positioning could lead to participants bumping into each other or going with the flow, the original author stated that he had actually used equidistant placement of the boxes on opposite sides of the exit door. Hence, this is what we used.

Table 1. Frequencies (and percentages) of choice for advertised and non-advertised pen in the liked and disliked music conditions, in the original study and the current three replications

	Liked music		Disliked music	
	Advertised pen	Non-advertised pen	Advertised pen	Non-advertised pen
Original Experiment (Gorn, 1982, Experiment 1)	74 (79%)	20 (21%)	30 (30%)	71 (70%)
		$N = 195$; $\chi^2(1) = 47.01$, $p < .001$, $\varphi = .49$		
Experiment 1 (exact replication)	34 (48%)	37 (52%)	39 (54%)	33 (46%)
		$N = 143$; $\chi^2(1) = .56$, $p = .45$, $\varphi = -.06$		
Experiment 2 (exact replication with additions)	38 (43%)	50 (57%)	38 (53%)	34 (47%)
		$N = 160$; $\chi^2(1) = 1.46$, $p = .23$, $\varphi = -.10$		
Experiment 3 (replication with updated music)	21 (57%)	16 (43%)	8 (23%)	27 (77%)
		$N = 72$; $\chi^2(1) = 8.59$, $p = .003$, $\varphi = .35$		

Notes. Depicted chi-square test results are confirmatory analysis equal to original study, excluding participants with deviant music evaluation.

Results

Two participants were removed from the sample because we could not reliably assess pen choice (form code and self-reported choice were incongruent). Analyses were conducted on the remaining 158 participants ($M_{age} = 20.28$, $SD = 1.48$; 37 male, 121 female). Music evaluations for "Summer Nights" were more positive ($M = 4.18$, $SD = 0.53$) than for the Indian music ($M = 2.27$, $SD = 0.67$; $F(1, 157) = 384.96$, 95% $CI_{diff} = 1.72–2.10$, $p < .001$, $\eta^2 = .71$), indicating that the music manipulation was successful.

Confirmatory Analyses

First, we replicated the original study's analysis by excluding participants who either (somewhat) disliked the liked music (evaluation below 3) or liked the disliked music (evaluation above 3), leaving 143 participants ($M_{age} = 20.26$, $SD = 1.43$; 35 male, 108 female; $N_{liked\_music} = 71$ (35 white pen, 36 blue pen), $N_{disliked\_music} = 72$ (33 white pen, 39 blue pen). Actual power of this sample to detect the original effect of $\varphi = .49$ is 1.00, power to detect the point null effect of $\varphi = .25$ is .85. Sensitivity analysis shows that the sample provides .95 power to detect a $\varphi = .30$ effect size, and .8 power to detect $\varphi = .23$.

A chi-square test of "advertised" versus "non-advertised" pen choice against "liked" versus "disliked" music showed no effect of music on pen choice, $\chi^2(1) = .56$, $p = .45$, $\varphi = -.06$, logOR = −.25, 95% CI = −.91 to .41). For cell frequencies, see Table 1. Because the 95% CI included 0, the main hypothesis was rejected. The obtained logOR is significantly lower than the original 2.17 (with 95% CI = 1.52–2.82). The obtained CI did not include the logOR of 1.04 associated with the point null effect of $\varphi = .25$. Based on the former criterion (e.g., Asendorpf et al., 2013), the current replication failed; based on the latter it should be regarded a conclusive failure (cf. Simonsohn, 2013).

Testing the main hypothesis with all 158 participants included ($M_{age} = 20.28$, $SD = 1.48$; 37 male, 121 female; $N_{liked\_music} = 73$ [35 white pen, 38 blue pen], $N_{disliked\_music} = 85$ [41 white, 44 blue]) showed no effect of music on pen choice, $\chi^2(1) = .21$, $p = .65$, $\varphi = -.04$, logOR = −.15, 95% CI = −.77 to .48; for liked music the ratio between advertised versus non-advertised pen choice was 36 versus 37; for disliked music 45 versus 40. As the 95% CI includes 0, the hypothesis was rejected.

Exploratory Analyses

Only 3.8% of participants mentioned music as a reason for choice, and 66.5% mentioned color preference. These results emulate those obtained in the original study. In describing the goal of the study, 46.2% of participants mentioned "influencing pen choice," indicating that many inferred pen choice was the main outcome variable. Hypothesis awareness was marginally higher for the liked music condition ($M = 2.08$, $SD = 1.68$) then for the disliked music condition ($M = 1.56$, $SD = 1.62$; $F(1, 157) = 3.87$, 95% $CI_{diff} = -.02$ to 1.04, $p = .05$, $\eta^2 = .02$), suggesting that demand artifacts could be more prominent in the former. However, logistic regression showed no effect of hypothesis awareness on choosing the "hypothesized" pen (OR = .93, 95% CI = .77–1.12, $p = .44$), indicating that hypothesis awareness did not transfer systematically into compliant or contravening pen choice. Thus, the current failure to replicate cannot be attributed to systematic biases in choice behavior of differentially hypothesis-aware participants. More exploratory results are reported in the supplementary materials.

Discussion

The current experiment showed no effect of music on pen choice. Because the 95% CI_{OR} included 0, and did not

include the originally obtained OR, nor an OR associated with the point null effect, the current results amount to a (conclusive) replication failure. Although some prior studies suggested that hypothesis awareness and resulting demand artifacts may have amplified or reduced the original study's effects, our results shows no systematic relation between hypothesis awareness and choice behavior.

Experiment 2

Compared to Experiment 1, two differences applied: (1) to avoid demand artifacts resulting from the presence of an involved researcher the experimenter team was naive; (2) to further explore possible demand artifacts, a more extensive post-experimental inquiry was conducted.

Participants

Participants were 195 second year BA Communication students, recruited through attendance of a persuasion class taught at a Dutch university. Participants were unaware of the purpose of the study, and did not receive credits or money for participation – they received credits for completing a complementary (unplanned) post-experimental questionnaire 2 weeks later.

Materials, Procedure, and Measures

Materials were the same as in Experiment 1. A male professional actor, posing as a researcher from another department, conducted the experiment with four assistants. All received a thorough briefing on the study's procedures, but were left naive regarding its purposes and hypotheses. Procedures were equal to Experiment 1's except for the inclusion of a more extensive post-experimental inquiry.

The planned post-experimental questionnaire contained questions (adapted from Allen & Madden, 1985) pertaining to demand artifacts (see supplementary materials). Two weeks later, 116 of the 195 participants filled out a secondary (not planned in preregistered proposal) online post-experimental questionnaire focusing on possible variations in experimental procedures, credibility of cover story, additional reasons for choice, and demand artifacts (see supplementary materials). Participants were debriefed in the subsequent class. Other measures were the same as in Experiment 1; music evaluation's α was .93.

Results

Five participants were excluded for having attended the course (featuring the original study) previously, leaving 190 participants. "Summer Nights" was evaluated more positively (M = 3.72, SD = .75) than the Indian music (M = 2.11, SD = .76; $F(1, 189) = 210.47$, 95% CI_{diff} = 1.39–1.83, $p < .001$, $\eta^2 = .53$), indicating a successful music manipulation.

Confirmatory Analyses

After excluding participants with an evaluation below 3 for the Grease song, and above 3 for the Indian music, 160 participants remained (M_{age} = 21.17, SD = 2.03; 42 male, 118 female; N_{liked_music} = 88 [51 white pen, 37 blue pen], $N_{disliked\_music}$ = 72 [28 white, 44 blue]; actual power to detect φ = .49: 1.00; power to detect the point null φ = .25: .89; .95 sensitivity: φ = .28; .8 sensitivity: φ = .22). A chi-square test showed no effect of music on pen choice, $\chi^2(1) = 1.46$, $p = .23$, $\varphi = -.10$, logOR = −.39, 95% CI = −1.01 to .24). See Table 1 for cell frequencies. The 95% CI included 0, the obtained logOR is not included in the original CI, and the obtained CI did not include the point null logOR of 1.04. Therefore, the current replication should be regarded a conclusive failure (Simonsohn, 2013).

Testing the main hypothesis on all 190 participants (M_{age} = 21.21, SD = 1.97; 55 male, 134 female, 1 n/a; N_{liked_music} = 110 [61 white pen, 49 blue pen], $N_{disliked\_music}$ = 80 [48 white, 32 blue]), similarly showed no effect of music on pen choice, $\chi^2(1) = 1.17$, $p = .28$, $\varphi = -.08$, logOR = −.32, 95% CI = −.90 to .26); liked music 49 (advertised pen) versus 61 (non-advertised pen); disliked music 42 versus 38. As the 95% CI includes 0, the hypothesis was rejected.

Exploratory Analyses

Again, few participants (3.7%) mentioned music as a reason for pen choice, and many (55.8%) mentioned color preference; 34.2% mentioned "influencing pen choice" as study goal. Contrasting to Experiment 1, hypothesis awareness did not differ between liked and disliked music conditions, $F(1, 184) = .70$, 95% CI_{diff} = −.22 to .41, $p = .42$, $\eta^2 = .00$. No effect of hypothesis awareness on choosing the "hypothesized" pen was found (OR = .94, 95% CI = .75–1.18, $p = .59$), indicating no systematic compliant or contravening behavior in hypothesis guessers.

In the unplanned secondary post-experimental inquiry, 74.1% of participants reported choosing the pen for own reasons; 7.4% indicated to have complied with the perceived study goal, whereas 18.8% contravened. Logistic regression shows that hypothesis awareness (determined from first post-experimental questionnaire) predicts these latter two behaviors combined (OR = 1.68, 95% CI = 1.21–2.34, $p = .002$); within the participants reacting on perceived study goals, hypothesis awareness elicited contravention rather than compliance (OR = 2.76, 95% CI = 1.17–6.55, $p = .02$). These results indicate that hypothesis awareness induces goal-contravening behaviors rather than goal-compliant behaviors. See supplementary materials for further exploratory analyses.

Discussion

The second experiment also did not show an effect of music on pen choice. The 95% CI included 0, and did not include the originally obtained effect size, nor the point null effect size. Thus, like Experiment 1, the current replication should be regarded a conclusive failure. The extensive post-experimental questionnaires showed that hypothesis awareness yields contravention to, rather than compliance with, perceived research goals. This makes sense: only by choosing the opposing pen participants can demonstrate they "outsmarted" the experimenters. Our results corroborate analyses by Shimp et al. (1991, 1993), who showed it was unlikely that demand artifacts caused the original findings (Gorn, 1982).

Experiment 3

Possibly, both direct replications described above failed because of the reuse of the original, over 30 years old, musical stimuli. Exposure to outdated music might elicit cognitive reflection on the experimental situation, and induce a state of involvement that could impede associative learning (Gorn, 1982, Experiment 2). Reflection might also enhance hypothesis awareness, in turn eliciting reactive responses (as seen in Experiment 2). Alternative to both prior experiments, Experiment 3 uses contemporary music.

Participants

Participants were 93 first year BA Communication students, recruited through attendance of an introduction on communication taught at a Dutch university. They were unaware of the purpose of the study, and did not receive credits or money for participation.

Materials, Procedure, and Measures

Because previous research attributed the original findings to the musical selections' differences in familiarity, lyrics, cultural origin, genre, tempo, and instrumentation (Kellaris & Cox, 1989), our aim was to select music similar on all these characteristics, and differing only in elicited affect. The pretest ($N = 47$) described above (see supplementary materials) also tested six contemporary pop songs against poor but professionally produced renditions by cover artists. Based on this pretest, we selected two renditions of the Rihanna song "We found love" as liked and disliked music. Both versions featured female singers and the same tempo, song sequence, and lyrics. Mean evaluations ($M = 5.60$ vs. 2.48) differed significantly, $F(1, 46) = 196.38$, 95% $CI_{diff} = 2.67–3.57$, $p < .001$, $\eta^2 = .81$.

The procedure of Experiment 3 emulated Experiment 2, omitting the elaborate post-experimental inquiry. Measures emulated Experiment 1; music evaluation's α was .87.

Results

Two participants were excluded because they were also enrolled in the second year BA class where Experiment 2 took place. Analyses were conducted on the remaining 91 participants. Music evaluation for the liked music (Rihanna) was more positive ($M = 3.73$, $SD = .77$) than for the disliked music (cover artist; $M = 2.19$, $SD = .77$; $F(1, 90) = 90.89$, 95% $CI_{diff} = 1.23–1.86$, $p < .001$, $\eta^2 = .51$), indicating the music manipulation was successful.

Confirmatory Analyses

After excluding participants with an evaluation below 3 for the liked music, and above 3 for the disliked music, 72 participants remained ($M_{age} = 19.26$, $SD = 2.13$; 17 male, 55 female; $N_{liked\_music} = 37$ [15 white pen, 22 blue pen], $N_{disliked\_music} = 35$ [17 white, 22 blue]; actual power to detect $\varphi = .49$: .99; power to detect the point null $\varphi = .25$: .56; .95 sensitivity: $\varphi = .42$; .8 sensitivity: $\varphi = .33$). This time, the chi-square test analyzing advertised versus non-advertised pen choice against liked versus disliked music showed a significant effect, $\chi^2(1) = 8.59$, $p = .003$, $\varphi = .35$, logOR = 1.49, 95% CI = .47–2.51. Cell frequencies were in the hypothesized direction (see Table 1). Because the 95% CI did not include 0, the main hypothesis was accepted. However, the obtained logOR of 1.49 is significantly lower than the original 2.17 (with 95% CI = 1.52–2.82). Employing this criterion (e.g., Asendorpf et al., 2013) the replication failed, even though the main hypothesis was accepted. Note that Simonsohn (2013) argued against considering replications failed when obtained effect sizes differ from the original. Instead he suggests considering replications that establish a significant effect in the hypothesized direction, and not significantly smaller than the point null effect, as successful (Simonsohn, 2013). Given that we acquired relatively small (.56) power to detect the point null effect, we concordantly qualify the current findings as a somewhat unreliable replication success.

Testing the main hypothesis on all 91 participants ($M_{age} = 19.25$, $SD = 1.97$; 20 male, 71 female; $N_{liked\_music} = 45$ [16 white pen, 29 blue pen], $N_{disliked\_music} = 46$ [24 white, 22 blue]) showed similar results: liked music promoted advertised pen choice, $\chi^2(1) = 4.03$, $p = .045$, $\varphi = .21$, logOR = .87, 95% CI = .01–1.73; for liked music, advertised versus non-advertised pen choice was 23 versus 22; for disliked music 14 versus 32. As the 95% CI did not include 0, the hypothesis was accepted.

Exploratory Analyses

Exploratory analyses showed that the effects observed in the current study stem largely from one experimental group (disliked music/blue pen), where 20 out of 22 participants chose the white pen. Worried about experimental anomaly, we tested whether this group differed from other groups regarding reasons provided for pen choice (e.g., "one box was better accessible"; "I followed a friend") or hypothesis

awareness. We found no significant differences (see supplementary materials). Also, the research assistants on site reported no anomalies in their post-experimental assessment report.

Of the total sample, 5.5% mentioned music influence and 58.2% color preference as a reason for choice; 58.2% mentioned the central item "influencing pen choice" as study goal. Hypothesis awareness was similar for liked and disliked music, $F(1, 86) = .67$, 95% $CI_{diff} = -.74$ to .39, $p = .54$, $\eta^2 = .00$. Notably, hypothesis awareness negatively affected choosing the "hypothesized" pen (OR = .71, 95% CI = .50–1.00, $p = .05$). More aware participants tended to contravene study goals, indicating that the observed effects of music on pen choice cannot be attributed to compliant behavior of participants "in the know."

Discussion

The final experiment showed the hypothesized effect of music on pen choice. The obtained effect size was significantly smaller than the original, but exceeded the null point effect. Although these results qualify a "successful replication" cf. Simonsohn (2013), achieved power was fairly low, and observed effects originated mostly in one experimental group. Therefore, conclusions from Experiment 3 should be drawn cautiously. Observed effects cannot be attributed to demand artifacts – hypotheses-aware participants chose the "hypothesized" pen significantly less often.

Aggregated Results

Aggregating our data, and excluding participants with "deviant" musical taste, we found no effect of music on pen choice, $N = 375$; $\chi^2(1) = .00$, $p = .99$, $\varphi = .00$, logOR = .00, 95% CI = −.41 to .40. The 95% CI included 0, and did not include the original 2.17, nor the OR of 1.04 associated with a point null effect, nor the OR of 0.44 associated with Simonsohn's (2013) $\varphi_{33\%}$ criterion (for which our aggregated sample provides .57 power). In unison, the three experiments failed to replicate the original results.

Adding the reconstructed data (195 cases) from the original study to ours (thereby aggregating all known direct replications; $N = 570$) the original effect still holds up, $\chi^2(1) = 15.54$, $p < .001$, $\varphi = .17$, logOR = .67, 95% CI = .33–1.00, due to the strong effect obtained in the original study. However, the aggregated effect size is small (Cohen, 1988).

Conclusion

Five conclusions can be drawn: (1) two well-powered replications failed to reproduce the original effect (Gorn, 1982). If the reader accepts the proposed $\varphi = .25$ point null effect, both can be considered conclusive replication failures cf. Simonsohn (2013); (2) a smaller replication using updated and matching musical selections – sufficiently powered to reliably detect the original effect, but featuring one experimental group with extreme scores – reproduced the original findings, but with a significantly smaller effect size. All in all, we labeled it a somewhat unreliable successful replication; (3) in aggregate, our studies conclusively failed to replicate the original effect; (4) aggregated data from all four known direct replications (including the original study) still show an effect of music on pen choice, though with considerably smaller effect size; (5) hypothesis awareness tends to elicit contravening rather than compliant responses in participants, rendering it unlikely that the original results were due to demand artifacts, as previously implied (Allen & Madden, 1985; Kellaris & Cox, 1989).

If, as suggested by Experiments 1 and 2, musical conditioning effects on pen choice do not exist, more replications would be needed to fully neutralize the original effect. Notably, 2,207 additional cases would be needed to push the aggregated effect below significance level (assuming future replications would consistently produce null effects).

Alternatively, if, as implied by Experiment 3, the proposed effect does sometimes emerge, moderators or confounds may be at play. Note that the original musical stimuli differ on more characteristics than elicited affect alone: for example, familiarity, lyrics, cultural origin, genre, tempo, and instrumentation. It is possible that these specific differences either amplified effects in the original 1982 sample (Kellaris & Cox, 1989) or dampened effects in the present samples. Note that Experiment 3's stimuli were not only contemporary, but also matched on all above, possibly confounding, differences, which discounts them as alternative explanations of observed effects. In addition, the current results discount participants' hypothesis awareness as an alternative explanation of the observed effects.

Limitations

The current research has several limitations, some of which may inform future replicators. First, it is uncertain whether we fully reproduced the original musical materials in Experiments 1 and 2. We selected the song "Summer Nights" as liked music because it posed no lyrical confounds, in contrast to the other candidate Grease songs. Yet, possibly, the original study used the song "You're the one that I want" to advertise pens. If so, this would provide a compelling alternative explanation of the strong effects observed.

Second, it was impossible to fully reproduce the original experimental context, not only because times and locations were different, but also because in general classroom experiments are very susceptible to noise. Small procedural variations or disruptions may influence entire experimental conditions, and group dynamics may influence individuals' behaviors. One might even question whether the current experimental set-up is suited to reliably determine subtle conditioning effects. To advance knowledge on the potential

of single-exposure musical conditioning of consumer choices, future (conceptual) replicators might better employ individualized experiments.

Third, we did not achieve planned power for our experiments. This is problematic for Experiment 3, which had relatively low power and therefore was susceptible to chance findings. Possibly, the extreme scores observed in one experimental group were such a chance finding. Future replicators might prefer experimental set-ups in which sample sizes can be fully controlled. Had Experiment 3's sample size been as planned, we could have attributed it more weight.

Finally, our research's theoretical contribution is limited. By emulating the original, rather unconventional, experimental set-up, we did not advance much in answering whether consumer choice can be conditioned through single exposure to music. We did establish, however, that the original well-cited findings (Gorn, 1982) were not – to paraphrase our title – a one-hit wonder, as testified by Experiment 3. To determine whether the findings constitute a theoretical "breakthrough," however, much care should be taken to eliminate possible confounds, preferably in non-classroom conceptual replications.

Acknowledgments

We thank Gerry Gorn for his support and for providing us with many details about the original experiment that helped us replicate it as faithfully as possible. We also thank Thijs Brenk and Sander Wensink, as well as 539 anonymous students at VU Amsterdam and KU Leuven, who spent considerable time and effort in participating in pretests and experimental sessions. We report all data exclusions, manipulations, and measures, and how we determined our sample sizes. This research was funded with a replicator grant of $1980 from the Center for Open Science. The authors declare no conflict-of-interest with the content of this article. Designed research: I.V., A.B., C.B.; Performed research: A.B., C.B., T.S., I.V.; Coded data: C.B., T.S., I.V.; Analyzed data: I.V., C.B., A.B., T.S.; Wrote paper: I.V., A.B., C.B., T.S. All materials, data, descriptions of the procedure, and the preregistered design are available at http://osf.io/ietwv/, registered at http://osf.io/cbn4x/. The design was registered, prior to data collection, at http://osf.io/z6e8j.

References

Allen, C. T., & Madden, T. J. (1985). A closer look at classical conditioning. *Journal of Consumer Research, 12*, 301–315.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119.

Bierley, C., McSweeney, F. K., & Vannieuwkerk, R. (1985). Classical conditioning of preferences for stimuli. *Journal of Consumer Research, 12*, 316–323.

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *British Medical Journal, 320*, 1468.

Cialdini, R. B. (2001). *Influence: Science and Practice* (4th ed.). Boston, MD: Allyn & Bacon.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1–11.

Fennis, B. M., & Stroebe, W. (2010). *The psychology of advertising*. New York: Psychology Press.

Gorn, G. J. (1982). The effects of music in advertising on choice behavior: A classical conditioning approach. *Journal of Marketing, 46*, 94. doi: 10.2307/1251163

Groenland, A. G. E., & Schoormans, J. P. L. (1994). Comparing mood-induction and affective conditioning as mechanisms influencing product evaluation and product choice. *Psychology and Marketing, 11*, 183–197.

Kellaris, J. J., & Cox, A. D. (1989). The effects of background music in advertising: A reassessment. *Journal of Consumer Research, 16*, 113. doi: 10.1086/209199

Peck, J., & Childers, T. L. (2008). Effects of Sensory Factors on Consumer Behavior. In C. P. Haugtvedt, P. M. Herr, & F. R. Kardes (Eds.), *Handbook of consumer psychology* (pp. 193–219). New York, NY: Erlbaum.

Saad, G. (2007). *The evolutionary bases of consumption*. Mahwah, NJ: Erlbaum.

Shimp, T. A., Hyatt, E. M., & Snyder, D. J. (1991). A critical appraisal of demand artifacts in consumer research. *Journal of Consumer Research, 18*, 273–283.

Shimp, T. A., Hyatt, E. M., & Snyder, D. J. (1993). A critique of Darley and Lim's "alternative perspective". *Journal of Consumer Research, 20*, 496–501.

Simonsohn, U. (2013). *Evaluating replication results.* Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2259879

Stuart, E. W., Shimp, T. A., & Engle, R. W. (1987). Classical conditioning of consumer attitudes: Four experiments in an advertising context. *Journal of Consumer Research, 14*, 334–351.

Received February 28, 2013
Accepted December 21, 2013
Published online May 19, 2014

Ivar Vermeulen

Faculty of Social Sciences
Communication Science
VU University Amsterdam
De Boelelaan 1081
1081 HV Amsterdam
The Netherlands
Tel. +31 20 598-9190
E-mail i.e.vermeulen@vu.nl